
**Information and documentation —
Thesauri and interoperability with other
vocabularies —**

**Part 2:
Interoperability with other vocabularies**

*Information et documentation — Thésaurus et interopérabilité avec
d'autres vocabulaires —*

Partie 2: Interopérabilité avec d'autres vocabulaires



STANDARDSISO.COM : Click to view the full PDF of ISO 25964-2:2013



COPYRIGHT PROTECTED DOCUMENT

© ISO 2013

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword	v
Introduction.....	vi
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
4 Symbols, abbreviations and other conventions	15
5 Objectives and identification	16
5.1 Objectives of interoperability and mapping	16
5.2 Identifying the elements to be mapped.....	16
6 Structural models for mapping across vocabularies	16
6.1 General	16
6.2 Model 1: Structural unity	17
6.3 Model 2: Direct-linked	17
6.4 Model 3: Hub structure	17
6.5 Selective mapping	19
6.6 Choosing among the options.....	20
7 Types of mapping.....	20
7.1 General	20
8 Equivalence mappings.....	21
8.1 General	21
8.2 Simple equivalence	21
8.3 Compound equivalence	22
9 Hierarchical mappings	25
10 Associative mappings.....	25
11 Exact, inexact and partial equivalence.....	26
11.1 General	26
11.2 Exact equivalence	26
11.3 Inexact equivalence	27
11.4 Partial equivalence	27
12 Use of mappings in information retrieval	28
12.1 General	28
12.2 Human mediation	28
12.3 Examples of mapping implementation.....	29
12.4 Conclusions and recommendations	30
13 Handling pre-coordination	31
13.1 General	31
13.2 Mapping between a thesaurus and a scheme with context-dependent captions	32
13.3 Mapping more complex classes	37
14 Techniques for identifying candidate mappings	38
14.1 General procedure.....	38
14.2 Computer assisted direct matching	40
14.3 Co-occurrence mapping	40
14.4 Other methods	40
15 Managing the data	41
15.1 Types of data to be recorded	41
15.2 Storing the data	43
15.3 Maintaining the mappings data.....	43

16	Display of mapped vocabularies	45
16.1	General.....	45
16.2	Single record display	46
16.3	Complete displays based on one of the vocabularies.....	48
16.4	Crosswalks	49
17	Classification schemes	50
17.1	Key characteristics and background.....	50
17.2	Semantic components and relationships compared with thesaurus components	53
17.3	Recommendations for mapping between a thesaurus and a classification scheme	55
18	Classification schemes used for records management	56
18.1	Key characteristics and background.....	56
18.2	Semantic components and relationships compared with thesaurus components	57
18.3	Recommendations for interoperability with a thesaurus	58
19	Taxonomies	59
19.1	Key characteristics and background.....	59
19.2	Types of taxonomies	60
19.3	Semantic components and relationships	61
19.4	Mapping between a thesaurus and a taxonomy.....	63
20	Subject heading schemes.....	67
20.1	Key characteristics and background.....	67
20.2	Semantic components and relationships	68
20.3	Mapping subject headings to and from thesaurus concepts	69
21	Ontologies	72
21.1	Key characteristics and background.....	72
21.2	Semantic components and relationships	73
21.3	Structural comparison between thesauri and ontologies	76
21.4	Interoperability with thesauri.....	77
22	Terminologies	78
22.1	Key characteristics and background.....	78
22.2	Semantic components and relationships compared with those of a thesaurus	79
22.3	Interoperability with thesauri.....	81
23	Name authority lists	81
23.1	Key characteristics and background.....	81
23.2	Semantic components and relationships	83
23.3	Mapping between a thesaurus and a name authority list.....	85
24	Synonym rings	87
24.1	Key characteristics and background.....	87
24.2	Semantic components and relationships	88
24.3	Interoperability with thesauri.....	89
Annex A (informative)	Management of terminological data in support of interoperability	90
Bibliography		93
Index.....		96

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 25964-2 was prepared by Technical Committee ISO/TC 46, *Information and documentation*, Subcommittee SC 9, *Identification and description*.

ISO 25964-1:2011, Clauses 1 to 13, correspond broadly to the content of ISO 2788:1986 and ISO 5964:1985. The remaining clauses of ISO 25964-1, and the whole of this part of ISO 25964, cover new ground not previously covered in any International Standard.

ISO 25964 consists of the following parts, under the general title *Information and documentation — Thesauri and interoperability with other vocabularies*:

- *Part 1: Thesauri for information retrieval*
- *Part 2: Interoperability with other vocabularies*

ISO 25964-1 covers the development and maintenance of thesauri, both monolingual and multilingual, including formats and protocols for data exchange.

This part of ISO 25964 covers interoperability between different thesauri and with other types of structured vocabulary.

ISO 25964 covers development and maintenance of thesauri rather than how to use them in indexing, which is covered by ISO 999.

Introduction

The ability to identify and locate relevant information among vast collections and other resources is a major and pressing challenge today, hence the need for semantic interoperability. To support this need, a variety of web services and other tools are under active development, including (but not limited to):

- registries of vocabularies and metadata schemas,
- repositories of vocabularies and metadata schemas, and
- crosswalk services.

While registries and repositories are outside the scope, this International Standard provides guidance fundamental to the establishment of crosswalk services. Typically these rely on mapping between different schemas and vocabularies. Accordingly, inter-vocabulary mapping will be the principal focus of this part of ISO 25964.

The foundation for discussion of interoperability is laid in ISO 25964-1, which describes the key characteristics of thesauri and provides guidance on best practice. Unfortunately, however, a comparable International Standard does not exist for the other types of vocabulary with which a thesaurus needs to interoperate. For this reason, this second part of ISO 25964 provides some elementary description of other vocabularies such as classification schemes, subject heading schemes, etc., before making recommendations on mapping between these and thesauri. It does not provide guidance on the construction of vocabularies other than thesauri.

Clauses 1 to 12 and Clauses 14 to 16 of this part of ISO 25964 deal with the principles and practicalities of interoperability, especially mapping, that apply to most vocabularies and especially thesauri. Clause 13 provides additional guidance for handling the pre-coordinated classes that occur in classification schemes and other vocabularies using the classification approach.

Clauses 17 to 24 are each dedicated to a different type of vocabulary. First priority is given to vocabularies that are commonly used for classifying or indexing resources, namely classification schemes (including those used for records management), taxonomies, subject heading schemes and name authority lists. Notwithstanding their different purposes, terminologies, ontologies and synonym rings are also included. Each clause provides a brief informative description of the vocabulary's key characteristics, contrasting its semantic components with those of a thesaurus, so as to provide context for the interoperability requirements and recommendations that follow.

Information and documentation — Thesauri and interoperability with other vocabularies —

Part 2: Interoperability with other vocabularies

1 Scope

This part of ISO 25964 is applicable to thesauri and other types of vocabulary that are commonly used for information retrieval. It describes, compares and contrasts the elements and features of these vocabularies that are implicated when interoperability is needed. It gives recommendations for the establishment and maintenance of mappings between multiple thesauri, or between thesauri and other types of vocabularies.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 25964-1:2011, *Information and documentation — Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

3.1

array

group of sibling **concepts** (3.17)

EXAMPLE

In the following, the sibling concepts outerwear and underwear form an array within the concept clothing.

clothing

outerwear

overcoats

underwear

3.2

associative relationship

relationship between a pair of **concepts** (3.17) that are not related hierarchically but share a strong semantic connection

3.3

broader term

preferred term (3.62) representing a **concept** (3.17) that is broader than the one in question

NOTE The scope of the narrower concept falls completely within the scope of the broader. The relationship between the two is commonly indicated with the tag BT. For more explanation see ISO 25964-1:2011, 10.2.1.

3.4

caption

text label of the **class** (3.10) represented by a **notation** (3.54) in a **classification scheme** (3.12)

NOTE A caption is read in conjunction with its hierarchical context. It does not need to be as complete or as self-contained as a scope note or even a preferred term in a thesaurus. Captions are sometimes known as class headings or class names.

3.5

category

concept (3.17) or group of similar or related **concepts** (3.17) used as a division or subdivision in a **taxonomy** (3.83)

NOTE 1 In a classification scheme, such a group is normally called a class.

NOTE 2 This definition of category should not be confused with “fundamental categories”, discussed in ISO 25964-1:2011 (for example in Clause 12) in the context of faceted classification.

3.6

category label

text label representing a **category** (3.5) in a **taxonomy** (3.83)

NOTE The category label in a taxonomy is comparable to the caption in a classification scheme, and although a category is often referred to as a node, a category label should not be confused with a node label.

3.7

chain index

index to a scheme that represents compound **concepts** (3.17) by a string of **pre-coordinated** (3.60) **terms** (3.84), such as a **classification scheme** (3.12), in which entries are generated by successive left truncation of strings of **terms** (3.84) representing complex **concepts** (3.17)

NOTE See the example in 17.2.4, Figure 12.

3.8

characteristic of division

attribute by which a **concept** (3.17) can be subdivided into an **array** (3.1) of narrower **concepts** (3.17) each having a distinct value of that attribute

cf. **node label** (3.52)

EXAMPLE

In the following, age group is the characteristic of division applied to the concept of people.

people
(people by age group)
children
youths
adults

3.9

citation order

order in which **preferred terms** (3.62) or **notations** (3.54) are combined in a **pre-coordinate** (3.60) **indexing** (3.36) system or a **classification scheme** (3.12) to form strings representing complex **concepts** (3.17)

3.10**class**

concept (3.17) or group of similar or related **concepts** (3.17) used as a division or subdivision in a **classification scheme** (3.12)

NOTE Classes are the basic units of which a classification scheme is constructed. In a taxonomy, although this is a type of classification scheme, they are generally known as categories (see 3.5). Occasionally they are described as “nodes”, although they should not be confused with node labels. The term “class” has a different meaning in the context of ontologies. See 21.2.2.

3.11**classification****classifying**

activity involving the components of grouping similar or related things together, separating dissimilar or unrelated things, and arranging the resulting **classes** (3.10) in a logical and helpful sequence

3.12**classification scheme**

schedule (3.67) of **concepts** (3.17) and **pre-coordinated** (3.60) combinations of **concepts** (3.17), arranged by **classification** (3.11)

NOTE A classification scheme often also includes an index.

3.13**collection**

set of information resources that can or could be accessed by a **structured vocabulary** (3.74), whether the items in the set are collected in one place or distributed over a network

3.14**compound equivalence**

relationship between **terms** (3.84) or **mapping** (3.41) between **concepts** (3.17) in which one **term** (3.84) or **concept** (3.17) in one context is represented by two or more **terms** (3.84) or **concepts** (3.17) in another

3.15**compound heading**

pre-coordinated heading

subject heading (3.76) formed by **pre-coordination** (3.60) of more than one **term** (3.84) into a string

EXAMPLE

The single terms Buddhism, Mass media and Religious aspects may be combined to form the compound heading
Mass media — Religious aspects — Buddhism.

3.16**compound term**

term (3.84) that can be split morphologically into separate components

EXAMPLES

In English:
“copper mines” can be split into “copper” and “mines”
“lawnmowers” can be split into “lawns” and “mowers”
In French:
“mine de cuivre” can be split into “mine” and “cuivre”
“biodiversité” can be split into “biologie” and “diversité”

NOTE Compound terms can be multi-word terms, or can consist of only one word.

3.17

concept

unit of thought

NOTE Concepts can be expressed in a variety of different ways. They exist in the mind as abstract entities independent of terms used to express them. They range from the very simple, e.g. "child", to the very complex, e.g. "child protection legislation".

3.18

concept group

group of **concepts** (3.17) selected by some specified criterion, such as relevance to a particular subject area

3.19

controlled vocabulary

prescribed list of **terms** (3.84), headings or codes, each representing a **concept** (3.17)

NOTE Controlled vocabularies are designed for applications in which it is useful to identify each concept with one consistent label, for example when classifying documents, indexing them and/or searching them. Thesauri, subject heading schemes and name authority lists are examples of controlled vocabularies.

cf. **terminology** (3.85)

3.20

cross-language equivalence

equivalence relationship (3.28) between **terms** (3.84) representing the same **concept** (3.17) in different languages

3.21

crosswalk

table of **mappings** (3.41) between the **concepts** (3.17) in two or more **structured vocabularies** (3.74)

NOTE Crosswalks support the ability of search engines to search effectively across heterogeneous databases.

3.22

data model

abstract model that describes how data is represented and used

NOTE The data model in ISO 25964-1 provides a generic definition of thesaurus structure and semantics. It can be used as the basis for defining a database model or an exchange format for thesauri.

3.23

differentiated mapping

mapping (3.40) methodology that aims to distinguish between different types and qualities of **mapping** (3.41)

NOTE Types of mapping that can be distinguished include equivalence, associative and hierarchical; equivalence can be further subdivided into simple or compound, and the degree of equivalence can be marked to support further distinctions at the time of use.

cf. **undifferentiated mapping** (3.88)

3.24

document

any resource that can be **classified** (3.11) or indexed in order that the data or information in it may be retrieved

NOTE This definition refers not only to written and printed materials in paper or microform versions (for example, conventional books, journals, diagrams, maps), but also to non-printed media such as machine-readable and digitized records, internet and intranet resources, films, sound recordings, people and organizations as knowledge resources, buildings, sites, monuments, three-dimensional objects or realia; and collections of such items or parts of such items.

3.25**entry term**

lead-in term

term (3.84) provided in a **controlled vocabulary** (3.19), not for direct use in **metadata** (3.45), but for the purpose of guiding the user to another **term** (3.84) which may be used as a **category label** (3.6), **subject heading** (3.76) or **preferred term** (3.62)

NOTE Entry terms occurring in a thesaurus are generally known as non-preferred terms.

3.26**enumerative classification scheme**

classification scheme (3.12) in which all the **classes** (3.10) available are explicitly listed in the **schedules** (3.67)

cf. **synthetic classification scheme** (3.80)

3.27**equivalence mapping**

mapping (3.41) that states that the **concept** (3.17) in the **target vocabulary** (3.82) is considered identical in scope to the **concept** (3.17) in the **source vocabulary** (3.72)

cf. **equivalence relationship** (3.28)

3.28**equivalence relationship**

relationship between two **terms** (3.84) in a **thesaurus** (3.86) that both represent the same **concept** (3.17)

NOTE In ordinary discourse, terms that are quasi-synonyms may represent slightly different concepts. After inclusion in the thesaurus, however, the equivalence relationship clarifies that both are regarded as representing the same concept. When two or more such terms are in the same language within a monolingual or multilingual thesaurus, one of them is designated a preferred term and the other(s) as non-preferred term(s); when two or more such terms are in the different languages of a multilingual thesaurus, each of them may be a preferred term in its own language respectively, and the relationship is known as cross-language equivalence.

3.29**exchange format**

machine-readable format for representing information that is intended to facilitate exchange of the information between different applications

NOTE The exchange format for a thesaurus often uses a markup language based on a standard such as XML (Extensible Markup Language), and is based on a data model for thesauri. While the data model provides a generic description of thesaurus structure and semantics, the exchange format expresses this in a formal language for the purpose of exchanging thesauri.

3.30**facet**

grouping of **concepts** (3.17) of the same inherent category

EXAMPLE 1

Animals, mice, daffodils and bacteria could all be members of a living organisms facet.

EXAMPLE 2

Digging, writing and cooking could all be members of an actions facet.

EXAMPLE 3

Paris, the United Kingdom and the Alps could all be members of a places facet.

NOTE Examples of high-level categories that can be used for grouping concepts into facets are: objects, materials, agents, actions, places and times.

cf. **node label** (3.52)

3.31

faceted classification scheme

classification scheme in which **concepts** (3.17) are analysed into their constituent **facets** (3.30)

NOTE Schedules are compiled for each facet, and terms or notations from these are then combined according to prescribed rules to express a complex concept. Some complex concepts are often enumerated explicitly in the schedules; others can be synthesized by the user.

3.32

hierarchical relationship

relationship between a pair of **concepts** (3.17) of which one has a scope falling completely within the scope of the other

cf. **broader term** (3.3), **narrower term** (3.51)

NOTE Several different types of hierarchical relationship exist. For more explanation see ISO 25964-1:2011, 10.2.

3.33

homograph

one of two or more words that are written in the same way, but have different meanings

EXAMPLE

In English:
The word “bank” could refer to a financial institution or the side of a river.
In French:
The word “avocat” could refer to a lawyer or to a fruit.

NOTE Homographs are sometimes referred to as homonyms, although the latter term applies more broadly, as it also includes pairs of terms such as “weights” and “waits” in English or “mer” and “mère” in French, which sound the same although they are spelt differently.

3.34

identifier

set of symbols, usually alphanumeric, designating a **concept** (3.17) or a **term** (3.84) or another entity for purposes of unique identification within a determined context or resource, especially in a computer system or network

NOTE A notation is sometimes used as an identifier.

3.35

index term

term (3.84) assigned to a **document** (3.24) in the process of **indexing** (3.36)

NOTE Sometimes index terms are referred to as indexing terms, as keywords or as tags, but the latter terms have other meanings too. Preferred terms from a thesaurus are very often used as index terms.

3.36**indexing**

intellectual analysis of the **subject** (3.75) matter of a **document** (3.24) to identify the **concepts** (3.17) represented in it, and allocation of the corresponding **index terms** (3.35) to allow the information to be retrieved

NOTE The term “subject indexing” is often used for this activity, but as ISO 25964 does not deal with the indexing of other elements such as authors or dates, “indexing” is sufficient. Indexing can be carried out by human users or by automated agents.

3.37**information retrieval**

all the techniques and processes used to identify **documents** (3.24) relevant to an information need, from a **collection** (3.13) or network of information resources

NOTE Selection and inclusion of items in the collection are included in this definition; likewise browsing and other forms of information seeking.

3.38**interoperability**

ability of two or more systems or components to exchange information and to use the information that has been exchanged

NOTE Vocabularies can support interoperability by including mappings to other vocabularies, by presenting data in standard formats and by using systems that support common computer protocols.

3.39**map, verb**

establish relationships between the **concepts** (3.17) of one vocabulary and those of another

3.40**mapping, gerund (verbal noun)**

process of establishing relationships between the **concepts** (3.17) of one vocabulary and those of another

3.41**mapping, noun**

⟨product of mapping process⟩ relationship between a **concept** (3.17) in one vocabulary and one or more **concepts** (3.17) in another

NOTE 1 A mapping generally has a direction, as discussed in Clause 6.

NOTE 2 Exceptionally, a mapping may involve a combination of two or more target vocabularies, where one or more of them act(s) as qualifier to the other (see 8.3.4).

3.42**mapping cluster**

coordinated set of **mappings** (3.41) between the **concepts** (3.17) of three or more vocabularies

NOTE See examples of mapping clusters in 6.3 and 6.4. A mapping cluster is generally maintained and published with a particular publishing or application objective. For example, a cluster of mappings between four different thesauri might be maintained so that the user of any one of them can easily search document collections indexed with any of the four.

3.43**markup**

annotations or other type of encoding embedded in text, in conformity with a **markup language** (3.44)

3.44

markup language

set of encoding conventions that can be used to provide instructions for the interpretation of a text, by the use of annotations embedded in the text itself

NOTE The interpretation often concerns issues such as content, structure or rendering of the text. Widely used examples include HTML (Hypertext Markup Language), which is largely concerned with presentation, and XML (Extensible Markup Language), which addresses the structure of text.

3.45

metadata

data that identify attributes of a **document** (3.24) typically used to support functions such as location, discovery, documentation, evaluation and/or selection

NOTE Preferred terms or notations selected during the indexing process are commonly applied as metadata values.

3.46

microthesaurus

designated subset of a **thesaurus** (3.86) that is capable of functioning as a complete **thesaurus** (3.86)

3.47

monohierarchical structure

hierarchical arrangement of **concepts** (3.17), in a **thesaurus** (3.86) or **classification scheme** (3.12), in which each **concept** (3.17) can have only one broader **concept** (3.17) at the level immediately above

cf. **polyhierarchical structure** (3.58)

EXAMPLE

In a monohierarchical structure, the concept of pianos cannot be listed under keyboard instruments as well as under stringed instruments; a choice has to be made of one of these concepts to determine its placing.

3.48

multilingual thesaurus

thesaurus (3.86) in which **terms** (3.84) and relational structures are available in two or more natural languages

3.49

multi-word term

term (3.84) consisting of more than one word

cf. **compound term** (3.16)

EXAMPLE

cost benefit analysis

3.50

name authority list

controlled vocabulary (3.19) for use in naming particular entities consistently

NOTE The entities in question are unique individuals, such as Benjamin Disraeli, Kilimanjaro or the Bayeux Tapestry, rather than classes such as politicians, mountains or embroideries. A name authority list may also be known as a name authority file. In this part of ISO 25964, a name authority list is sometimes referred to simply as an authority list.

3.51**narrower term**

preferred term (3.62) representing a **concept** (3.17) which is narrower than the one in question

NOTE The scope of the narrower concept falls completely within the scope of the broader. The relationship between the two is commonly indicated with the tag NT. For more explanation see ISO 25964-1:2011, 10.2.1.

3.52**node label**

label inserted into a hierarchical or **classified** (3.11) display to show how the **terms** (3.84) have been arranged

NOTE A node label is neither a preferred term nor a non-preferred term. It contains one of two different types of information: either a) the name of a facet to which following terms belong; or b) the attribute or characteristic of division by which an array of sibling concepts has been sorted or grouped. See examples in ISO 25964-1:2011, Clause 11.

3.53**non-preferred term**

non-descriptor

term (3.84) that is not assigned to **documents** (3.24) but is provided as an entry point in a **thesaurus** (3.86) or index

cf. **entry term** (3.25)

EXAMPLE

<i>hounds</i> USE dogs

NOTE A non-preferred term is followed by a reference to the appropriate preferred term or preferred terms.

3.54**notation**

class code

class number

classmark

set of symbols representing a **concept** (3.17) or **class** (3.10) in a **structured vocabulary** (3.74), especially a **classification scheme** (3.12)

EXAMPLES

Notation	Source vocabulary	Concept
07.04.4	ILO Thesaurus	Fishery policy and development
622.342 2	Dewey Decimal Classification	gold mining
373.3.016:51	Universal Decimal Classification	mathematics curriculum in primary schools
SBS XEJ B	Bliss Bibliographic Classification	endangered species law
H40-H42	International Statistical Classification of Diseases and Related Health Problems	Glaucoma

NOTE Notation is sometimes used to sort and/or locate concepts in a pre-determined systematic order and, optionally, to display how the components of complex concepts have been structured and grouped. A notation can provide the link between alphabetical and systematic lists in a thesaurus. In the context of classification schemes, "concepts" are often known as "subjects", especially when they are complex as in the examples above.

3.55

one-to-many mapping

mapping (3.41) in which a single **concept** (3.17) in one vocabulary is **mapped** (3.39) to a combination of two or more **concepts** (3.17) in another vocabulary

NOTE This situation is different from one in which a concept has two or more independent mappings to concepts in the other vocabulary.

3.56

one-to-one mapping

mapping (3.41) in which a single **concept** (3.17) in one vocabulary is **mapped** (3.39) to a single **concept** (3.17) in another vocabulary

NOTE 1 The terms or notations used to label the concepts in the two vocabularies may or may not be identical.

NOTE 2 It is possible for a single concept to have two or more one-to-one mappings, if these are independent of each other.

3.57

ontology

a formal, explicit specification of a shared conceptualization

NOTE This definition is attributable to Studer et al.^[31], extending an earlier definition by Gruber^[18], and is adopted in this part of ISO 25964 because it is widely accepted in the ontology development community. An ontology typically includes definitions of concepts and specified relationships between them, set out in a formal way so that a machine can use them for reasoning. This definition excludes thesauri, classification schemes and other structured vocabularies described in this part of ISO 25964, even though these are sometimes described as “lightweight ontologies”.

3.58

polyhierarchical structure

hierarchical arrangement of **concepts** (3.17), in a **thesaurus** (3.86) or **classification scheme** (3.12), in which each **concept** (3.17) can have more than one broader **concept** (3.17)

cf. **monohierarchical structure** (3.47)

EXAMPLE

In a polyhierarchical structure, organs (musical instruments) could be listed under keyboard instruments as well as under wind instruments.

NOTE In a polyhierarchical structure, a single concept can occur at more than one place in the hierarchical structure of the thesaurus. Its attributes and relationships, and specifically its narrower and related terms, are the same wherever it occurs.

3.59

post-coordination

combination of **preferred terms** (3.62) of a **controlled vocabulary** (3.19) at the time of searching

cf. **pre-coordination** (3.60)

EXAMPLE

The post-coordinated search expression “microwaves AND radiation” can be used to retrieve documents on microwave radiation, when these have been indexed under the separate terms “microwaves” and “radiation” rather than a compound term.

3.60**pre-coordination**

combination of **concepts** (3.17), **classes** (3.10) or **terms** (3.84) of a **controlled vocabulary** (3.19) at the time of its construction or at the time of using it for **indexing** (3.36) or **classification** (3.11)

cf. **post-coordination** (3.59)

EXAMPLE 1

The class “general theory”, when placed within the broader class “music”, refers only to the pre-coordinated subject “theory of music” and not to theory in general.

EXAMPLE 2

The pre-coordinated string “cardboard — recycling” might appear in a subject heading scheme, or if not enumerated there, it might be synthesized by an indexer when needed for a particular document.

3.61**precision**

measure of retrieval performance defined by R/T , where R is the number of relevant items retrieved and T is the total number of items retrieved from the same **collection** (3.13)

3.62**preferred term**

descriptor

term (3.84) used to represent a **concept** (3.17) when **indexing** (3.36)

cf. **non-preferred term** (3.53)

NOTE A preferred term is usually a noun or noun phrase.

3.63**protocol**

convention that defines the syntax, semantics and synchronization of the communication process between two computers in order to enable a particular service

3.64**quasi-synonym**

near-synonym

one of two or more **terms** (3.84) whose meanings are generally regarded as different in ordinary usage but which may be treated as labels for the same **concept** (3.17), in a given **controlled vocabulary** (3.19)

EXAMPLES

diseases, disorders
earthquakes, earth tremors

3.65**recall**

measure of retrieval performance defined by R/N , where R is the number of relevant items retrieved and N is the total number of relevant items in the **collection** (3.13)

3.66

related term

preferred term (3.62) representing a **concept** (3.17) which has an **associative relationship** (3.2) with the one in question

NOTE The relationship between related terms is commonly indicated with the tag RT. For more explanation, see ISO 25964-1:2011, 10.3.

3.67

schedule

terms (3.84), **notations** (3.54), **captions** (3.4), cross-references and **scope notes** (3.68) set out to exhibit the content and structure of a **structured vocabulary** (3.74)

3.68

scope note

note that defines or clarifies the semantic boundaries of a **concept** (3.17) as it is used in the **structured vocabulary** (3.74)

NOTE A term used to label a concept can have several meanings in normal usage. A scope note is used to restrict the concept to only one of those meanings, and where necessary refers to other concepts that are included or excluded from the scope of the concept being clarified.

3.69

search term

term (3.84) forming all or part of a search query

NOTE In the context of ISO 25964, search terms are usually drawn from a controlled vocabulary.

3.70

search thesaurus

vocabulary intended to assist searching even though it has not been used to **index** (3.36) the **documents** (3.24) being searched

NOTE Search thesauri are designed to facilitate choice of terms and/or expansion of search expressions to include terms for broader, narrower or related concepts, as well as synonyms. Optionally, a thesaurus complying with this part of ISO 25964 can be used as a search thesaurus.

3.71

source language

language serving as a starting point in translation or in a search for **term** (3.84) equivalents

3.72

source vocabulary

vocabulary that serves as a starting point when seeking a corresponding **term** (3.84) or **concept** (3.17) in another vocabulary

3.73

specificity

capability of a **structured vocabulary** (3.74) to express a **subject** (3.75) in depth and in detail

NOTE For more explanation see the discussion of specificity in ISO 25964-1:2011, 8.4, and other places.

3.74

structured vocabulary

organized set of **terms** (3.84), headings or codes representing **concepts** (3.17) and their inter-relationships, which can be used to support **information retrieval** (3.37)

NOTE A structured vocabulary can also be used for other purposes. In the context of information retrieval, the vocabulary needs to be accompanied by rules for how to apply the terms. Various types of structured vocabulary are discussed in this part of ISO 25964, including classification schemes, subject heading schemes, etc.

3.75**subject**

concept (3.17) or combination of **concepts** (3.17) dealt with in a **document** (3.24) or occurring in discourse

3.76**subject heading**

term (3.84) or pre-coordinated string of **terms** (3.84), drawn from a **subject heading scheme** (3.77)

3.77**subject heading scheme**

subject heading language

subject heading list

SHL

structured vocabulary (3.74) comprising **terms** (3.84) available for subject **indexing** (3.36), plus rules for combining them into pre-coordinated strings of **terms** (3.84) where necessary

3.78**synonym**

one of two or more **terms** (3.84) denoting the same **concept** (3.17)

EXAMPLES

In English: guarantees, warranties heart attack, myocardial infarction HIV, human immunodeficiency virus
In French: schiste, phyllade VIH, virus de l'immunodéficience humaine crise cardiaque, infarctus du myocarde

NOTE Abbreviations and their full forms can be treated as synonyms.

3.79**synonym ring**

set of synonymous or almost synonymous **terms** (3.84), any of which can be used to refer to a particular **concept** (3.17)

EXAMPLE

stream; river; brook; beck; burn

3.80**synthetic classification scheme**

classification scheme (3.12) in which users can synthesize **notation** (3.54) for complex **classes** (3.10) from lists of simpler **classes** (3.10)

cf. **enumerative classification scheme** (3.26)

3.81**target language**

language providing a translation or an equivalent for a **term** (3.84) existing in a **source language** (3.71)

3.82**target vocabulary**

vocabulary in which a **term** (3.84) or **concept** (3.17) is sought corresponding to an existing **term** (3.84) or **concept** (3.17) in a **source vocabulary** (3.72)

3.83

taxonomy

scheme of **categories** (3.5) and subcategories that can be used to sort and otherwise organize items of knowledge or information

NOTE Taxonomies range from the very simple to the very complex. In the simplest taxonomies the categories are not necessarily divided into subcategories, while in the complex ones multiple hierarchical levels can be found. Other features too can be present, such as all the thesaurus features described in ISO 25964-1, and/or the features commonly found in classification schemes. Outside of this International Standard, the term is often used loosely to refer to any type of structured vocabulary.

3.84

term

word or phrase used to label a **concept** (3.17)

EXAMPLES

schools
school uniform
costs of schooling
teaching

NOTE Thesaurus terms can be either preferred terms or non-preferred terms.

3.85

terminology

set of designations belonging to one special language

[ISO 1087-1:2000, definition 3.5.1]

NOTE The term “special language” is defined in ISO 1087 to mean a “language used in a subject field and characterized by the use of specific linguistic means of expression”; while a “designation” is defined therein as a “representation of a concept by a sign which denotes it”.

3.86

thesaurus

controlled (3.19) and **structured vocabulary** (3.74) in which **concepts** (3.17) are represented by **terms** (3.84), organized so that relationships between **concepts** (3.17) are made explicit, and **preferred terms** (3.62) are accompanied by lead-in entries for **synonyms** (3.78) or **quasi-synonyms** (3.64)

NOTE The purpose of a thesaurus is to guide both the indexer and the searcher to select the same preferred term or combination of preferred terms to represent a given subject. For this reason a thesaurus is optimized for human navigability and terminological coverage of a domain.

3.87

top term

preferred term (3.62) representing a **concept** (3.17) that has no broader **concept** (3.17) in the **thesaurus** (3.86)

NOTE Sometimes indicated by the abbreviation TT.

3.88

undifferentiated mapping

mapping (3.40) methodology that does not distinguish between different types of **mapping** (3.40), nor indicate different degrees of equivalence

cf. **differentiated mapping** (3.23)

3.89

vocabulary control

management of a vocabulary in order to disambiguate and constrain the form of the **terms** (3.84) and limit the number of **concepts** (3.17) and **terms** (3.84) available for **indexing** (3.36)

NOTE Control is achieved by distinguishing between homographs, so that each one has just one meaning, and by picking out from a set of synonyms or quasi-synonyms the one which is to be preferred for use in indexing. The purpose of these restrictions is to increase the likelihood of indexers and searchers choosing the same term to label a particular concept.

4 Symbols, abbreviations and other conventions

For the purposes of this part of ISO 25964, the symbols, abbreviations and conventions in ISO 25964-1:2011, Clause 3 apply. Additional tags and symbols are set out in Table 1. The alpha-2 language codes from ISO 639-1 are also used in this International Standard.

In some examples, labels such as “VOC1” and “VOC2” are used to signify “Vocabulary 1”, “Vocabulary 2”, etc.

To indicate the direction of a mapping statement, one option is to use an arrow explicitly. Alternatively, in languages that are customarily read from left to right, the source vocabulary concept should be shown on the left, followed by the appropriate tag(s) from Table 1, followed by the target concept(s).

Table 1 — Additional abbreviations and symbols used in mappings

Symbol	Meaning				
EQ	Equivalence; the term that follows the tag is the preferred term in a target vocabulary that is closest in meaning to the preferred term preceding the tag, from the source vocabulary.				
=	This symbol (an equals sign, Unicode character U+003D) should be used in conjunction with a mapping abbreviation to indicate that the mapping is exact. Specifically, “=EQ” signifies exact equivalence.				
~	This symbol (a tilde, Unicode character U+007E) should be used in conjunction with a mapping abbreviation to indicate that the mapping is inexact. Specifically, “~EQ” signifies inexact equivalence.				
BM	Broader mapping; the term that follows the tag represents a concept having a wider meaning.				
NM	Narrower mapping; the term that follows the tag refers to a concept with a more specific meaning.				
RM	Related mapping; the term that follows the tag labels an associated concept, but is not a synonym, a quasi-synonym, a broader term or a narrower term.				
	This symbol, (a vertical bar or “pipe line”, Unicode character U+007C), is shown between two or more preferred terms from a target vocabulary, the scopes of which, in sum, best cover the scope of a broader preferred term in the source vocabulary. Each of the target concepts represents part of the scope of the source concept. When converting index terms, all the target vocabulary preferred terms should be applied to a record indexed with the source vocabulary term. When converting search statements, the target vocabulary preferred terms should be combined with Boolean OR. EXAMPLE <table> <tr> <td><u>Source vocabulary</u></td><td><u>Target vocabulary</u></td></tr> <tr> <td>Livestock</td><td>sheep cattle pigs poultry</td></tr> </table>	<u>Source vocabulary</u>	<u>Target vocabulary</u>	Livestock	sheep cattle pigs poultry
<u>Source vocabulary</u>	<u>Target vocabulary</u>				
Livestock	sheep cattle pigs poultry				
+	This symbol (a plus sign, Unicode character U+002B) is shown between two or more preferred terms from a target vocabulary that are used in conjunction to represent a compound concept in the source vocabulary. Each of the target concepts represents an aspect of the source concept. When converting index terms, all the target vocabulary preferred terms should be applied to a record indexed with the source vocabulary term. When converting search statements, the target vocabulary preferred terms should be combined with Boolean AND. EXAMPLE <table> <tr> <td><u>Source vocabulary</u></td><td><u>Target vocabulary</u></td></tr> <tr> <td>women executives</td><td>women + executives</td></tr> </table>	<u>Source vocabulary</u>	<u>Target vocabulary</u>	women executives	women + executives
<u>Source vocabulary</u>	<u>Target vocabulary</u>				
women executives	women + executives				

5 Objectives and identification

5.1 Objectives of interoperability and mapping

During information retrieval, the principal aim of interoperability between vocabularies is to enable an expression formulated using one vocabulary to be converted to (or supplemented by) a corresponding expression in one or more other vocabularies. This applies whether or not the vocabularies use the same natural language. The expression in question may be either a search query or part of the metadata associated with a document. In both cases, mapping is the key step. If each of the concepts in Vocabulary A has been mapped to the corresponding concept(s) in Vocabulary B, it becomes possible to interchange (or augment) the terms or identifiers representing the concept in each of the vocabularies. At the working level, interoperability is enabled by establishing inter-concept mappings, particularly equivalence, for which guidance and recommendations are given in Clauses 7 to 13.

NOTE Another aspect of interoperability is enhancement of the vocabulary tools through activities such as merging multiple vocabularies or using portions of one to extend another. This International Standard should not be interpreted as restricting novel forms of interoperability that could emerge.

5.2 Identifying the elements to be mapped

For all types of vocabulary, mappings aim to show relationships between concepts in different vocabularies, although the concepts are generally known as “classes” in the context of a classification scheme, and “categories” in the case of a taxonomy. This International Standard often uses the term “concept” to take in all of these.

Table 2 shows the principal elements that are used to represent concepts in different vocabulary types. Statements of mappings should use these elements when intended for human readers. But when mappings are intended for manipulation by computer, the concepts should instead be represented by unique, persistent identifiers.

Table 2 — Human-readable elements used to represent concepts in mapping statements

Vocabulary type	Concepts represented by:
Thesaurus	Preferred terms
Classification scheme	Notations
Taxonomy	Category labels or notations (see Note)
Subject heading scheme	Headings
Name authority list	Names
Ontology	Labels
Terminology	Terms or other types of designation
NOTE Different styles of concept representation are used by different taxonomies. Where there is no notation or unique identifier and the category labels are not unique, it is usually necessary to spell out the whole hierarchical path to specify a given concept uniquely.	

6 Structural models for mapping across vocabularies

6.1 General

This clause deals with the overall models within which equivalence and other mappings should be managed. Three basic models are described in 6.2, 6.3 and 6.4, and an alternative approach in 6.5. Management of the mappings data is discussed in Clause 15.

6.2 Model 1: Structural unity

In the structural unity model, all of the participating vocabularies share exactly the same structure of hierarchical and associative relationships between concepts. The structure may be represented or expressed in any number of different languages, notations or coding systems. The data model in ISO 25964-1:2011, Clause 15, illustrates structural unity as applied to a symmetrical multilingual thesaurus and discussed in ISO 25964-1, 9.1. The structural simplicity of this model makes it both feasible and desirable to manage all the concepts, terms, notations, captions and relationships between them within one system.

NOTE Model 1 does not usually involve mapping, but is included for comprehensiveness and for comparison with the other models. Some applications call for Model 1 to be applied in combination with other models.

6.3 Model 2: Direct-linked

The direct-linked model addresses linkages between two or more vocabularies that do not share the same structure. As well as differing in scope, language and structure, the vocabularies may include other types of vocabulary (classification scheme, name authority list, etc.) as well as one or more thesauri. Direct mappings should be established between the concepts of each vocabulary and those of each other vocabulary. This model may be extended to any number of vocabularies, by establishing direct mappings from each vocabulary to each other one. Any one of the vocabularies may then be used to search any collection indexed with any other one of the vocabularies.

Figure 1 illustrates the mappings that are needed to handle four vocabularies, using the direct-linked model. The boxes represent the concepts in the four vocabularies, respectively. The mappings are represented by arrows between the concepts. Each arrow has a direction. For simplicity, Figure 1 shows all the mappings as double-headed arrows, indicating that the mappings are intended to work in both directions. To describe this more accurately, each double-headed arrow represents a pair of mappings, one in each direction. Thus Figure 1 shows a total of 12 sets of mappings, presented in six pairs.

To reduce the amount of mapping work, sometimes mappings in one direction only are established, enabling conversion of index terms or search terms in one direction only. One-way mappings would be represented by single-headed arrows in the same basic model.

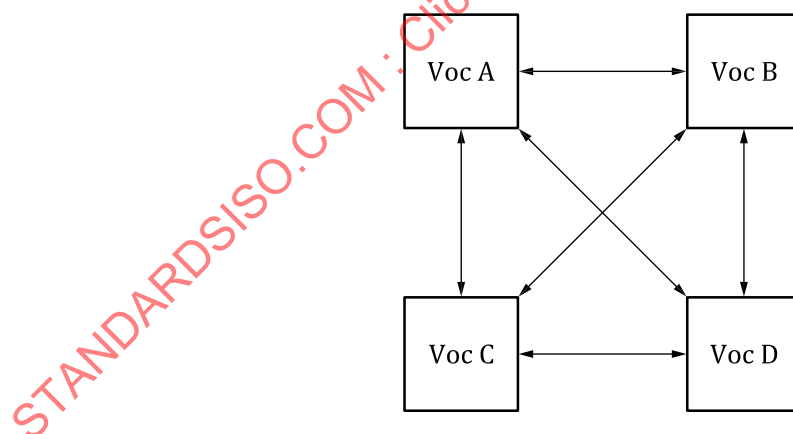


Figure 1 — Model 2 (direct-linked) as applied to four vocabularies

6.4 Model 3: Hub structure

Instead of establishing direct mappings between each pair of vocabularies, it is often convenient to designate one vocabulary as a “hub”, or comprehensive structure to which each of the other vocabularies is mapped. Figure 2 shows Model 3, in which vocabulary B serves as hub and the others as “satellites”.

Using this model, each concept in the hub vocabulary should be mapped to the corresponding concept(s) in the other vocabularies, and vice versa. In the same style as for Model 2, double-headed arrows indicate mappings in both directions.

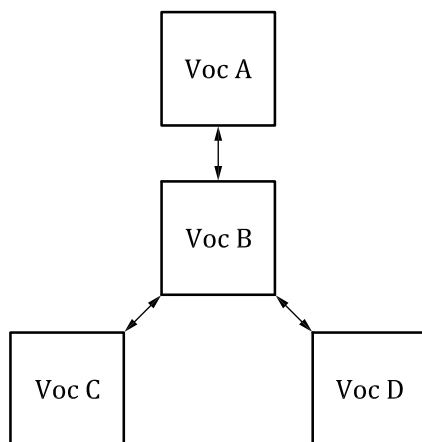


Figure 2 — Model 3 (hub model) as applied to four vocabularies

This model enables the hub vocabulary to be used for searching any of the resources indexed with any of the satellite vocabularies, either by conversion of search queries or by conversion of index terms. Similarly, an index term or a search term drawn from any of the satellite vocabularies can be converted to the corresponding term(s) in the hub vocabulary. A third possibility is to apply two conversions sequentially, for example from Voc A to Voc B, and then from Voc B to Voc D. When both steps involve exact, simple equivalence (see Clause 8 and 11.2), sequential mapping gives entirely acceptable results. But if the first step involves compound equivalence, or if two inexact mappings are applied in sequence, the quality of the result is unpredictable.

To achieve good quality mappings for the entire scope of the satellite vocabularies, the hub vocabulary needs to incorporate all the concepts present in all the satellites. When there is not a vocabulary of sufficient scope and specificity among those to be mapped, it may be necessary to construct a new hub vocabulary to meet the need. Alternatively, one of the original vocabularies may serve as hub if it is first extended to match the combined scope and depth of all the others.

For some applications, two-way mappings are not necessary. Figure 3 shows two variations on Model 3, in which mappings apply in one direction only.

- The variation on the left would allow a user of Vocabulary Q to search resources indexed with Vocabularies P, R or S. Alternatively, if the mappings in it were used to convert the metadata of resources indexed with Vocabulary Q, users of Vocabularies P, R and S would be enabled to search those resources.
- The variation on the right would allow users of Vocabularies W, Y and Z to search resources indexed with Vocabulary X. Alternatively, if the mappings in it were used to convert the metadata of resources indexed with Vocabularies W, Y or Z, users of Vocabulary X would be enabled to search those resources.

In neither case is it possible to convert terms from any one of the satellite vocabularies to any of the others.

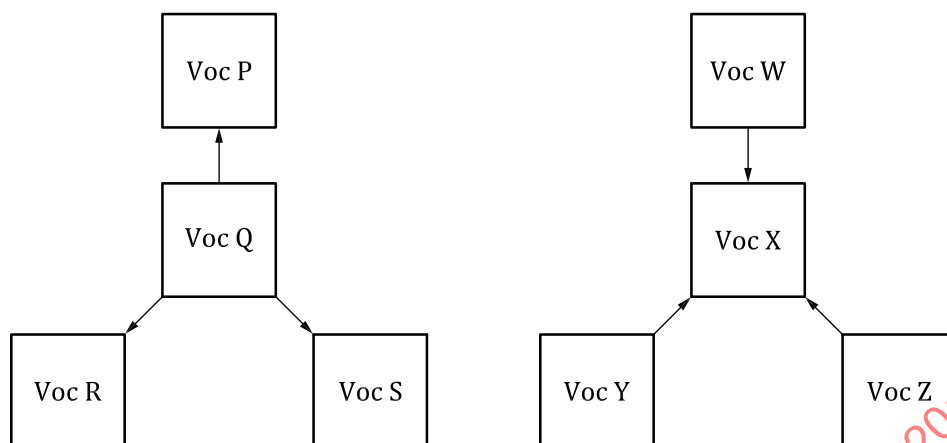


Figure 3 — Two variations on Model 3, with one-way mappings

6.5 Selective mapping

Models 1, 2 and 3 all require significant work to build and maintain. For some applications, however, it is unnecessary to map the vocabularies comprehensively. One alternative is to establish mappings only for the concepts that have been used or are likely to be used within the application in question. Two cases can be distinguished:

a) When overlap is small

If there are relatively few concepts common to two or more vocabularies, only a limited number of mappings can and should be established. See Figure 4.

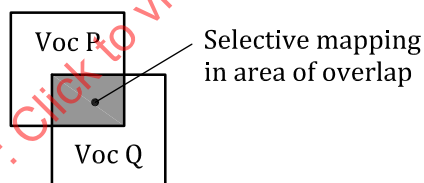


Figure 4 — Selective mapping where vocabulary overlap is small

b) Mapping from an index or catalogue

An approach that is often found convenient is to establish mappings for the headings (or terms or notations) that occur in an index or catalogue of the collection(s) involved in the application. Using this approach, the index or catalogue is treated as a sub-vocabulary derived from the full version of the vocabulary. This reduces the initial mapping effort, but can increase updating maintenance when changes are made in the collection. Figure 5 illustrates the approach when applied to a catalogue.

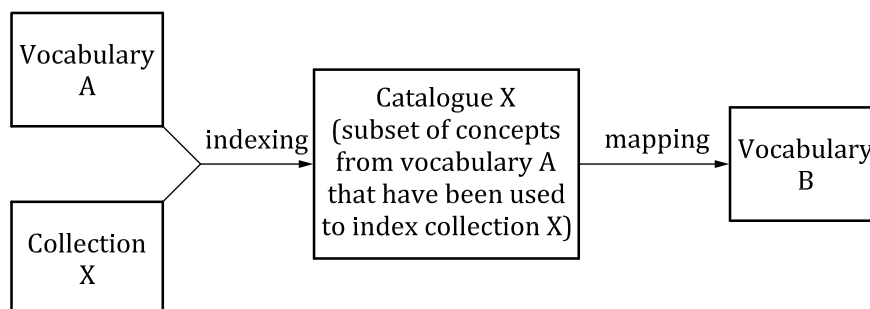


Figure 5 — Mapping between catalogue entries and a vocabulary

When a catalogue or index is used as the basis of mapping to more than one other vocabulary, it may be treated as the hub vocabulary in Model 3. Typically the mappings will enable conversion in one direction only, from the hub to the satellites.

6.6 Choosing among the options

The three structural models have been enumerated separately but in real applications combinations of them often occur and the boundaries might be blurred. At the start of any mapping project, it is important to clarify which model or combination of models will be used, with particular attention to the direction of the mappings.

In choosing among the options, the following factors should be borne in mind.

- a) The structural unity model (Model 1) is most applicable when a new thesaurus is being prepared for use in more than one language, as described in ISO 25964-1. The absence of an existing indexed collection makes it easier to give each language equal status.
- b) Models 2 and 3 are appropriate for the reconciliation of vocabularies that have been independently developed and/or have already been applied to collections. Model 2 is applicable particularly when there are only two or three such vocabularies.
- c) Model 2 is more powerful than Model 3 if direct interoperability between all pairs of vocabularies is needed because it avoids sequential mapping. However, direct interoperability between the satellite vocabularies is not always needed. And the maintenance costs for Model 2 tend to be higher.
- d) Between any pair of vocabularies, the mapping quality that can be achieved is best when the target vocabulary has equal specificity as well as the same breadth of coverage as the source vocabulary. (The mapping examples in 16.2 illustrate the difficulty when a target vocabulary is sparse in the subject area required.)
- e) As a consequence of d), a vocabulary chosen or built as hub should be rich in the required subject area(s), at all levels. (That is to say, it needs to have at least equal specificity as well as the same breadth of coverage, as all of the satellite vocabularies combined.)
- f) Among a group of vocabularies to be mapped, if one of them has significantly less specificity than the others, relatively few equivalence mappings will be possible. The more specific concepts in the other vocabularies can only be mapped hierarchically or associatively to the relatively generic concepts in the first. (See Clauses 7 to 10 for a discussion of the different types of mapping.)
- g) If a multilingual thesaurus includes some non-symmetrical structures, a hybrid approach may be adopted. Model 1 should be used wherever possible, but Model 2 may be adopted when it proves impossible to establish one-to-one equivalence across languages.
- h) Selective mapping based on an index or catalogue, as described in 6.5, has limited application. It may be less expensive to implement, but supports mappings for a particular application in one direction only. Since the index lacks some of the concepts and relationships in the vocabulary from which it was derived, the semantic structure of the vocabulary becomes incomplete and thus unavailable as a support for the mapping work or for the application of the mappings.

7 Types of mapping

7.1 General

When relationships are established across vocabularies, they are usually known as mappings. In principle, any type of relationship may be defined and applied as a mapping. In practice, the mappings likely to be useful are determined by the types of vocabulary to be linked. Between thesauri, the three main types of mapping to consider are equivalence, hierarchical and associative (exactly analogous to the relationships used internally within any thesaurus). Of these, equivalence is the type most commonly needed. Additional relationship types may be considered if another vocabulary type is involved, especially an ontology.

The three main mapping types, and their subdivisions, are described in Clauses 8 to 10. Clause 12 shows how they are used when mapping between thesauri. Clause 13 deals with the more complex case where one of the vocabularies to be mapped is a pre-coordinate scheme such as a classification scheme or subject heading scheme.

8 Equivalence mappings

8.1 General

Equivalence should be established when matching concepts are found in two or more different vocabularies. Unlike the establishment of equivalence between two terms in a single monolingual thesaurus, where one of them is designated the preferred term while the other is designated a non-preferred term (see ISO 25964-1:2011, Clause 9), in the context of cross-vocabulary equivalence the mapping is between concepts and there is no difference of status between the concepts or between the preferred terms or notations representing them. The tag “EQ” should be used to indicate equivalence in mapping statements.

Typically the vocabularies include different selections of concepts and develop these to different levels of specificity. Hence several different equivalence situations can arise, with different solutions as described in subclauses 8.2 to 11.

8.2 Simple equivalence

In the ideal situation, the target vocabulary contains a concept identical in scope to the concept in the source vocabulary. A one-to-one equivalence mapping may immediately be established between the concepts.

EXAMPLES

Vocabulary 1	Vocabulary 2
mobile phones	cell phones
aubergines	egg-plants

The mapping is conventionally expressed in the following format: “mobile phones EQ cell phones”.

If more than two vocabularies are under consideration, the mapping statement may need to distinguish between them. Clause 16 shows how this should be done in displays for human readers.

Very often the preferred terms for the matching concepts are identical, in which case a mapping statement such as: “dishwashers EQ dishwashers” is appropriate.

The converse does not always apply. Identical preferred terms in two different vocabularies should not be accepted as equivalents without checking the underlying concepts. For example, the preferred term “operations” could have different meanings in a military thesaurus and a medical thesaurus.

Even when the contexts are similar, there can be subtle differences of scope. For example, the concept labelled “teachers” in one vocabulary might include teaching staff at universities as well as schools, whereas another vocabulary might limit the scope of “teachers” to schoolteachers, and provide a separate concept labelled “lecturers” for teaching staff at universities. Similarly the term “public schools” has a very different meaning for American and British contexts, because the educational systems of these countries are differently organized. Where identical terms are found in different vocabularies, an equivalence mapping should be established only if the underlying concepts are judged equivalent.

NOTE Further examples of equivalence, including situations where different degrees of equivalence apply, are given in Clause 11.

8.3 Compound equivalence

8.3.1 General

A complex concept that is included in one vocabulary with just one preferred term may be represented in another vocabulary by a combination of two or more concepts/terms.

EXAMPLES

Vocabulary 1	Vocabulary 2
genetically modified wheat	genetic modification wheat
inland waterways	rivers canals
fossil fuels	coal petroleum oil natural gas

In such cases a compound equivalence mapping (also known as one-to-many equivalence) may be established between the concepts. The mapping generally applies in one direction only.

Two distinct types of compound equivalence are commonly found, known as “intersecting” and “cumulative” respectively.

8.3.2 Intersecting compound equivalence (EQ +)

The intersecting compound equivalence mapping type gets its name because it can be represented in set theory as the intersection of two or more sets. Figure 6 illustrates the first of the following examples.

EXAMPLES

Vocabulary 1	Vocabulary 2
women executives	women executives
genetically modified wheat	genetic modification wheat
abused children	children abuse
rail passenger safety	railways passengers safety

Intersecting compound equivalence mappings are not reversible. For example, if “women” and “executives” coexist in the metadata of a single document, the combination should not be mapped automatically to “women executives”, since the document might instead deal with relationships between male executives and other women.

The use of some examples of compound equivalence is illustrated in 12.3.

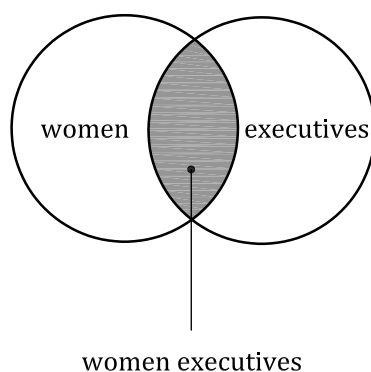


Figure 6 — The set of women who are also executives is the subset shown at the intersection of the two larger sets

A statement of intersecting compound equivalence should be conventionally expressed in the following format: “women executives EQ women + executives”

If the mappings will be used only for conversion of search expressions, the symbol “+” in mapping statements may be replaced by the Boolean operator “AND”. Boolean operators do not apply in the context of mapping index terms. See 12.3 for examples of how mappings are used in information retrieval.

8.3.3 Cumulative compound equivalence (EQ |)

The cumulative compound equivalence mapping type applies when a complex concept in one vocabulary corresponds to the sum of two or more concepts in another. Figure 7 shows its representation in set theory. Note that the simpler concepts do not necessarily overlap, but they will add up to the whole of the complex concept.

EXAMPLES

Vocabulary 1	Vocabulary 2
inland waterways	rivers canals
hosiery	stockings socks
woody plants	trees shrubs
fossil fuels	coal petroleum oil natural gas

Cumulative compound equivalence mappings should not be used in the reverse direction without checking whether another mapping would be more suitable. For example, in Vocabulary 1 the concept “inland waterways” might have narrower concepts “rivers” and “canals”. In this case, “rivers” and “canals” in Vocabulary 2 should be mapped to their simple equivalents, in preference to the broader concept.

The use of some examples of compound equivalence is illustrated in 12.3.

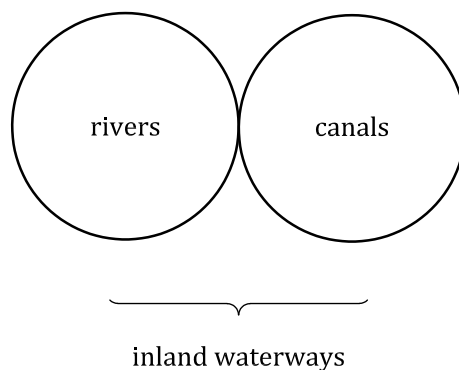


Figure 7 — The set of inland waterways is the superset composed of the set of rivers plus the set of canals

NOTE This example works best in contexts and regions where all the rivers are navigable, and there are no lakes or other bodies of water used as inland waterways. See 11.4 c) for advice on the situation where the narrower concepts available do not add up to the whole complex concept.

A statement of cumulative compound equivalence should be conventionally expressed in the following format: “inland waterways EQ rivers | canals”

If the mappings will be used only for conversion of search expressions, the symbol “|” in mapping statements may be replaced by the Boolean operator “OR”. Boolean operators do not apply in the context of mapping index terms.

8.3.4 Compound equivalence involving a combination of target vocabularies

A special case of compound equivalence can arise when the target vocabulary is inadequate to convey the concepts in a source vocabulary, unless it is accompanied by a “qualifier vocabulary”, that is to say a vocabulary composed of concepts that can be used to qualify the concepts in the main target vocabulary. In this unusual circumstance, a single concept in the source vocabulary may be mapped to a combination in which one concept is drawn from the main target vocabulary and the other(s) is/are selected from the qualifier vocabulary/ies.

EXAMPLE

A source vocabulary for languages includes a comprehensive listing of regional dialects and other sublanguages as well as main languages. In this vocabulary, the English spoken in Britain has preferred term “en-GB”, while that spoken in the United States of America has preferred term “en-US”.

For a particular application, mappings are required to a target vocabulary (VocL) which lists the main languages, but has no provision for dialects. It is accompanied by another “qualifying” target vocabulary (VocG) with a comprehensive listing of countries and regions. Compound equivalence mappings in the following style may be established:

en-GB EQ VocL:English + VocG:United Kingdom

en-US EQ VocL:English + VocG:United States of America

Combinations as described here should be used with caution, since they may have limited validity outside the particular application for which they are intended. The index term “United Kingdom”, for example, is normally used only for documents about this country, and could prove misleading if applied to situations where the country is not involved at all.

9 Hierarchical mappings

Hierarchical mapping may be established between concepts when one is clearly broader than the other. Just as for hierarchical relationships within a single thesaurus, described in ISO 25964-1:2011, 11.2, the relationship between the concepts can be generic or instantial. The whole-part situation may also justify establishment of a hierarchical mapping, in the limited circumstances described in ISO 25964-1:2011, 11.2.3.

EXAMPLES

Vocabulary 1	Vocabulary 2
school staff	schoolteachers
roads	streets
international organizations	United Nations
armies	battalions
United Kingdom	England

A hierarchical mapping from the narrower to the broader concept should be expressed in the following format: “streets BM roads”. The reciprocal hierarchical mapping, from the broader to the narrower concept, should be expressed in the following format: “roads NM streets”.

NOTE The tags BM and NM, short for Broader Mapping and Narrower Mapping respectively, are analogous to the BT and NT tags used within a single thesaurus. Use of a slightly different tag avoids confusion in displays that include cross-vocabulary mappings as well as internal relationships.

Optionally, it is possible to make a formal distinction between generic, instantial and whole-part hierarchical mappings, just as may be done for generic, instantial and whole-part hierarchical relationships within a single thesaurus. Instead of the tags BTG/NTG, BTI/NTI and BTP/NTP described in ISO 25964-1:2011, 10.2, the tags BMG/NMG, BMI/NMI and BMP/NMP should be used respectively, as illustrated in Table 3 below.

Table 3 — Mapping statements that distinguish between subtypes of hierarchical mapping

Subtype	Mapping statement example	Reciprocal example
Generic	rats BMG rodents	rodents NMG rats
Instancial	Paris BMI capital cities	capital cities NMI Paris
Whole-part	fingers BMP hands	hands NMP fingers

10 Associative mappings

The associative mapping type may be established between concepts when they do not qualify for equivalence or hierarchical mappings, but are semantically associated to such an extent that documents indexed with the one are likely to be relevant in a search for the other.

The dividing line between an associative mapping and inexact equivalence (see 11.3) is ill-defined and subjective, but may be made pragmatically according to the context in which the mappings are to be used (taking into account, for example, the interests of the users, the volume of relevant resources to be searched, and the capability of the retrieval and display systems to make fine distinctions).

EXAMPLES

Vocabulary 1	Vocabulary 2
discipline	punishment
e-learning	distance education
translation	interpreting

An associative mapping should be expressed in the following format: “discipline RM punishment”. In cases of associative mapping, a reciprocal mapping is valid, e.g. “punishment RM discipline” in this case.

NOTE The tag RM, short for Related Mapping, is analogous to the RT tag used within a single thesaurus. Use of a slightly different tag avoids confusion in displays that include cross-vocabulary mappings as well as internal relationships.

11 Exact, inexact and partial equivalence

11.1 General

Since the quality of a particular mapping is often less than perfect, it is optionally possible to apply a marker to an equivalence mapping, indicating the degree to which the mapping is universally applicable. Two degrees are distinguished: exact (see 11.2) and inexact (see 11.3), which should be marked using the symbols “=” and “~” respectively. The distinction should be applied only to cases of simple equivalence.

In cases of compound equivalence, exact/inexact markers would convey no extra information because all such mappings are to some extent inexact. Conversely, in the case of hierarchical mappings, only those which are indisputably “exact” should be accepted as hierarchical; the other candidates may be more appropriately characterized as associative mappings or inexact equivalents. Finally, in the case of associative mappings, these cover so many different possible types of relationship that it is not meaningful to designate them as exact or inexact.

11.2 Exact equivalence

An equivalence mapping between two concepts found in different vocabularies is said to be exact when the concepts can be used interchangeably across all the applications that can be envisaged for the mapping. If the mappings are intended for use in one or a small number of contexts only, each of these contexts should be considered before judging the equivalence exact. If the mappings are to be made more widely available, this judgement should be made across a wide range of information retrieval applications.

The symbol “=” should be used with “EQ” to mark exact equivalence in a mapping statement.

EXAMPLES

Vocabulary 1		Vocabulary 2
mad cow disease	=EQ	bovine spongiform encephalopathy
guinea pigs	=EQ	guineapigs

Exact equivalence mappings are by definition reversible; in other words, they may be applied as two-way mappings. Provided they have been established with care, they can also be applied sequentially without risk (see 6.4).

11.3 Inexact equivalence

Sometimes the most closely matching concepts in two or more vocabularies are not exactly the same. The problem is particularly acute when the vocabularies have emerged from different cultural communities (see also the discussion and examples in ISO 25964-1:2011, Clause 10). The following cases commonly arise.

- The concepts may be equivalent in some contexts but not others.
- The concepts may have overlapping scopes or small differences of connotation.
- When mapping between a classification scheme and a thesaurus, it is common to find a class with a caption that matches a preferred term, but checking shows that the two are not exactly equivalent. Reasons for this are discussed in more detail in Clause 13 and 17.2.

EXAMPLES

Vocabulary 1	Vocabulary 2
organizational structure	management structure
devolved government	devolution
lawns	turf
chairs	seats
astronomy in a thesaurus, as preferred term with "astrophysics" as a Related Term	astronomy in a classification scheme, as caption of a broad class that includes astrophysics
women executives	female managers
horticulture	gardening

An inexact equivalence mapping should be expressed in the following style: "lawns ~EQ turf".

Inexact equivalence may occur in cases of compound as well as simple equivalence, and should be expressed similarly.

When inexact equivalence occurs between overlapping concepts, the question can arise, "How great is the overlap?" This question is important in applications where a user needs to evaluate the likely benefit of following a mapping in order to retrieve more relevant information without retrieving too many irrelevant documents. Estimating the extent of overlap can also help with the ranking of search results. For such applications, it may be found useful to characterize the inexact equivalence as "major overlap" or "minor overlap".

11.4 Partial equivalence

Sometimes the only difference between closely matching concepts in different vocabularies is that one is slightly broader than the other. To put it another way, one of the concepts has a scope only partially equivalent to the other.

EXAMPLES

Vocabulary 1	Vocabulary 2
monarchs	kings
container plants	potted plants
aircraft	aeroplanes
school premises	school buildings

There is, however, no unique way of identifying partial equivalence. Instead, a difficult, rather subjective choice has to be made between the following options.

- a) Establish a hierarchical mapping (see Clause 9), e.g. “monarchs NM kings”.
- b) Establish an equivalence mapping, marking it as inexact if that distinction is being used, e.g. “container plants ~EQ potted plants”. This option should be chosen only if hierarchical mappings are not available in this application and option c) is not applicable.
- c) Establish a cumulative compound equivalence mapping. This is possible only when Vocabulary 2 also contains the concept(s) comprising the missing part of the concept in Vocabulary 1. For example, “aircraft EQ aeroplanes | helicopters”. If it is judged that an important part of the broader concept (e.g. hot air balloons, blimps and other miscellaneous aircraft) is likely to be missed, two or more separate hierarchical mappings should be established rather than one compound equivalence mapping (e.g. “aircraft NM aeroplanes”; “aircraft NM helicopters”).

12 Use of mappings in information retrieval

12.1 General

In the context of information retrieval, there are two main stages at which mappings between controlled vocabularies can be used: (a) as part of the indexing process or (b) at the time of search.

NOTE These are not the only situations in which mappings may be used, and the recommendations that follow should not be interpreted as excluding other applications.

- a) When mappings are used in the indexing process, index terms in the metadata of documents indexed with Vocabulary A are converted to (or augmented by) the corresponding terms from Vocabulary B. This can be done routinely while original indexing takes place or as a bulk conversion of a complete collection, supplemented by small updates each time the vocabularies are updated and/or new documents are added to the collection. The document collection may then be browsed or searched using Vocabulary B for the indefinite future, without need of additional mapping.
- b) When mappings are applied to search terms, the original metadata is not changed. To achieve the same capability (that of using Vocabulary B to search a collection indexed with Vocabulary A), the source and target vocabularies for mappings need to be established in the reverse direction. This enables search queries comprising terms from Vocabulary B to be converted to the corresponding terms from Vocabulary A. Thus the mappings are incorporated into the search process and need to be applied every time a search is done using Vocabulary B.

12.2 Human mediation

In some implementations, mappings are applied automatically, but in others the process is mediated by a person who checks the suitability of the mapping case by case. This is especially useful for mappings other than exact simple equivalence. In cases where no exact equivalent is available, the indexer makes a choice between any inexact and/or hierarchical mappings, selecting the one that best suits the document in question. Likewise a searcher can choose the mapping closest to his own information need.

EXAMPLE

Human mediation in a mapped search:

A user seeks information on “child protection”, which is a preferred term in Vocabulary A. But in Vocabulary B this term does not exist and instead the user is offered an inexact mapping to “safeguarding children” and/or an associative mapping to “child care”. The searcher can choose one or both of these terms, depending on which is closest to his need.

Human mediation should be used wherever feasible to achieve a higher retrieval performance and avoid false inferences from inexact equivalences.

12.3 Examples of mapping implementation

In the simplest approach to mapping, the objective is only to indicate pairs of concepts in different vocabularies that have some unspecified correspondence to each other. However, the precision and recall of retrieval applications can be greatly enhanced if mappings are differentiated using all the types of mapping described in Clauses 7 to 10, and if degrees of equivalence are marked as recommended in Clause 11.

To illustrate the power of differentiation, Table 4 shows examples of how the various types of mapping work in practice. For simplicity, it is assumed in all cases that the vocabularies involved are thesauri. (Clause 13 shows more complex examples, in which one of the vocabularies is a monohierarchical scheme such as a classification scheme.)

In the first column of Table 4 is an example of each of the main types of mapping described in Clauses 7 to 11.

In the second column, the first sentence for each example says how it should be implemented if the mappings are used to convert index terms from Vocabulary A to Vocabulary B. Automatic conversion is assumed, without any human mediation. Additional sentences comment on the likely performance of the mapped index terms when Vocabulary B is used for retrieval. The performance of the mapped terms is compared with the performance possible if the same documents had been indexed directly with Vocabulary B, rather than indirectly via mapping from Vocabulary A. The description of performance is inevitably somewhat hypothetical, because the index terms will be used to satisfy many different searches and the precision/recall characteristics of these will vary.

In the third column, the first sentence for each example says how it should be implemented if the mappings are used to convert search terms from Vocabulary A to Vocabulary B. Automatic conversion is assumed, without any human mediation. Additional sentences comment on the likely performance of the mapped search expression. The likely effects on recall and precision are compared with the performance possible if the document collection had been indexed with Vocabulary B.

**Table 4 — Comparison of mapping implementation
in the conversion of index terms and search terms**

Type of mapping/ Example	Applied to index terms	Applied to search terms
Exact, simple equivalence mad cow disease =EQ bovine spongiform encephalopathy	The index term “mad cow disease” is converted to “bovine spongiform encephalopathy”. No negative effects should occur, provided the controlled vocabulary is used effectively.	A search for “mad cow disease” is converted to “bovine spongiform encephalopathy”. Both precision and recall are as they would have been if the collection had been indexed with the target vocabulary.
Intersecting compound equivalence women executives EQ women + executives	Both broader terms are added to the metadata. This tends to increase recall in all searches involving the terms “women” or “executives”. However, users searching specifically for the concept of women executives will not easily separate the relevant items from others dealing both with executives of either sex, and with women in any role at all.	The search becomes “women AND executives”. Recall is unaffected; precision is reduced. This search will retrieve documents that concern executives of either sex as well as women, for example a document about male executives harassing women staff.

Table 4 (continued)

Type of mapping/ Example	Applied to index terms	Applied to search terms
Cumulative compound equivalence inland waterways EQ rivers canals	Both narrower terms are added to the metadata. Unless Vocabulary A also has the narrower terms “rivers” and “canals”, there will be a negative effect on precision (because a search for “canals” could retrieve items that were indexed with “inland waterways” despite dealing only with “rivers”).	The search becomes “rivers OR canals”. There is no negative effect on recall, but precision may be reduced if the user wants his search to apply at the generic level only.
Inexact simple equivalence potted plants ~EQ house plants	The index term “potted plants” is converted to “house plants”. For a search on house plants, precision is reduced, as it will retrieve some items dealing with potted plants situated outdoors. Recall too may be reduced, if Vocabulary A does not allow for the concept “house plants”, since items dealing with non-potted house plants will not have been indexed with either term.	The search term “potted plants” is converted to “house plants”. Precision is reduced, as the search may retrieve some items dealing with house plants that are in containers other than pots.
Hierarchical mapping (narrower to broader) streets BM roads	The index term “streets” is converted to “roads”. Any negative effect is likely to be small, provided the controlled vocabulary is used effectively.	The search term “streets” is converted to “roads”. Recall may be increased slightly and precision reduced, as the search has been broadened.
Hierarchical mapping (broader to narrower) roads NM streets	The index term “roads” is converted to “streets”. This will have a negative effect on the precision of searches for “streets” as they will now retrieve documents dealing with various types of roads (e.g. motorways and country lanes) that would not normally be considered streets.	The search term “roads” is converted to “streets”. Precision will not suffer but recall will be reduced, since only one type of road will be retrieved.
Associative mapping e-learning RM distance education	In this case the index term “e-learning” is converted to “distance education”. As a result, users searching for distance education can expect to retrieve some relevant documents, together with some that discuss other applications of e-learning.	In this case the search term “e-learning” is converted to “distance education”. A negative influence on both recall and precision can be expected.

12.4 Conclusions and recommendations

The examples in 12.3 show that some types of mapping are more useful than others, depending on the context of application and the relative specificities of the two vocabularies.

- Exact, simple equivalence is the only type that can be applied reversibly in any context, either to index terms or to search terms, without affecting the quality of retrieval.
- Cumulative compound equivalence can also be highly effective, provided it is applied in the direction intended and the combination of narrower concepts really adds up to the whole concept being mapped.

- Intersecting compound equivalence generally reduces the precision of searches.
- Neither form of compound equivalence can be used directly in the opposite direction. That is to say, the co-occurrence of two or more terms in a search statement or document metadata should not be used without other checks to infer their equivalence with one combined concept. However, the separate terms may be used to suggest hierarchical mappings in the reverse direction. For example, either “women” or “executives” or a combination of both would suggest a hierarchical mapping to the narrower concept of “women executives”. Conversely, either “rivers” or “canals” or a combination of both would suggest a hierarchical mapping to the broader concept of “inland waterways”.
- Hierarchical mappings tend to give more acceptable results in the direction from narrower to broader concept than in the reverse direction.
- Greater flexibility will be available in a particular application if the available mappings are not restricted to one per concept. Especially where a given concept has no exact equivalent, provision of inexact, broader, narrower and associative mappings is helpful in enabling selection of the best option for the context required.

At the start of any mapping project, the following choices need to be made:

- which overall model or combination of models to use (see Clause 6);
- how much to differentiate the mappings, in the following respects:
 - whether to distinguish between equivalence and other types of mapping such as hierarchical and associative (see Clauses 7 to 10);
 - whether to accept compound equivalence mappings (see 8.3);
 - whether to distinguish between exact and inexact equivalence (see Clause 11); and
 - whether to enable establishment of more than one mapping per concept;
- whether to implement the mappings at the point of indexing or at the point of search;
- in which direction to prepare the mappings, and whether two-way mapping is required; and
- whether and how to enable human mediation in the conversion process.

13 Handling pre-coordination

13.1 General

Some types of vocabulary provide for pre-coordinated concepts or headings, that is to say, complex concepts in which two or more simpler concepts have been combined in one. Sometimes the pre-coordinated headings are explicitly enumerated.

ENUMERATED EXAMPLES

a	344.032 Law of social welfare	<i>from DDC (Dewey Decimal Classification)</i>
b	Books — Africa — History	<i>from LCSH (Library of Congress Subject Headings)</i>

In addition to explicit enumeration, some vocabularies incorporate rules for synthesizing classes or headings when wanted, by combining simpler elements at the time of indexing.

SYNTHESIZED EXAMPLES

c	373.3.016:51 (Teaching mathematics in primary schools)	<i>This class is not enumerated in the UDC, but can be synthesized from a combination of 373.3 (Primary school. Elementary level) with 37.016 (Curriculum) and 51 (Mathematics).</i>
d	Pyramids — Egypt	<i>This heading is not enumerated in LCSH but can be synthesized from the enumerated headings Pyramids and Egypt.</i>

In a classification scheme, the pre-coordinated concepts are not always explicit in the accompanying caption. For example, in the scheme shown in Figure 8, the caption “institutions” appears in three different hierarchies, referring to three different classes. Within the broader class “Education”, this caption implies that subclass E100 refers to educational institutions. Within “Defence”, the same caption is used to indicate that D100 covers defence institutions.

E	Education	H	Health	D	Defence
E100	institutions	H100	institutions	D100	institutions
E200	activities	H200	activities	D200	activities
E210	training	H210	training	D210	training
	<i>medical training, see H210</i>	H220	preventive medicine	D220	operations
	<i>military training, see D210</i>	H230	operations		etc.
E211	teacher training	etc.	etc.		
	etc.				

Figure 8 — Fragments of a monohierarchical scheme with notation

The occurrence of pre-coordinated concepts, classes or headings creates an added challenge for interoperability. When exactly the same pre-coordinated concept occurs in two different vocabularies, a one-to-one mapping can and should be established. More usually the choice of which constituent concepts to combine varies from one vocabulary to another, and this leads to a frequent need for one-to-many mappings.

Pre-coordinated headings in the style of Enumerated Example b and Synthesized Example d, above, occur only with subject heading schemes. The way to handle them is described in Clause 20.

The style of pre-coordination illustrated in Figure 8 occurs not just in classification schemes, but more widely in any scheme with a monohierarchical structure, for example the schemes used in records management and other filing systems, and in many taxonomies. The implications for mapping between thesauri and any of these are discussed in 13.2.

The classes of a classification scheme are often intrinsically complex as in example a above. Further examples of this kind are discussed in 13.3.

13.2 Mapping between a thesaurus and a scheme with context-dependent captions

13.2.1 General

Mappings to or from a class or category in a monohierarchical scheme should treat the class/category as a pre-coordinated concept whose meaning can be established by inspecting all its superordinate and subordinate classes as well as any scope notes associated with it. Inspection of the caption alone is inadequate. For example, in Figure 8 “institutions” inadequately represents the class E100, which applies only to educational institutions.

Any of the mapping types — equivalence, hierarchical and associative (see Clauses 7 to 10) — may be used. If a single concept in the thesaurus and a class in the scheme match each other according to one of these types, a one-to-one mapping may be established.

Where there is no single concept to match a given class, especially one which involves pre-coordination, the class may be analysed to establish whether a one-to-many mapping can be found. This process is sometimes called parsing. Thus 373.3.016:51 in Enumerated Example c in 13.1 can be parsed into the simpler classes 373.3 (Primary school), 37.016 (a special auxiliary number for Curriculum, hence only .016 is used in synthesis) and 51 (Mathematics); matches for these simpler classes can then be sought in the thesaurus.

In a statement of the mapping, the pre-coordinated concept should be uniquely represented by its notation or by an identifier provided explicitly for this purpose. A caption alone should not be used to represent the class in a mapping statement.

To illustrate this section, the schema shown in Figure 8 will be used to provide examples. This schema could be drawn from a very simple classification scheme, a monohierarchical taxonomy or the scheme used by a records management system. The recommendations apply equally to all of these, and to any other vocabulary following a similar pre-coordinated monohierarchical model. As in Clause 12, mappings for implementation as part of the indexing process will be discussed separately from those implemented at the time of search.

13.2.2 Mappings for conversion of indexing/cataloguing entries when the thesaurus is the target vocabulary

The following guidelines should be applied.

- A mapping should be sought for each class in the classification scheme. The examples in Table 5 are based on the classes enumerated in Figure 8 for the Education and Health hierarchies only.
- For each class, the most closely matching thesaurus concept or combination of concepts should be sought by examining the preferred term, non-preferred terms and scope note of each candidate concept. In cases of doubt, the broader and narrower terms should also be checked as well as the way the preferred term has been used in indexing.
- If an exactly equivalent thesaurus concept is found, it is usually enough to establish one equivalence mapping. But if the best available match is inexact, broader, narrower or associative additional mappings may be established for all candidates that seem useful in the context(s) envisaged.
- Where exact simple equivalence is established, the mappings may be used to make fully automatic conversion of notations appearing in metadata to the corresponding index terms. In other cases, it is advisable to apply additional checks, such as human mediation in the process (see 12.2).

Table 5 — Sample mappings from a classification scheme to a thesaurus

Class	Found in the thesaurus	Mapping options
E	"education" as preferred term, with many narrower terms and related terms.	"E EQ education" The marker "=" or "~" may be applied, depending on whether the thesaurus term is intended to be used in the same sense as the class E.
H	No entry for "health" or "human health" but preferred terms "public health" and "animal health" are found.	"H EQ public health" The marker "=" or "~" may be applied, depending on whether the thesaurus term is intended to be used in the same sense as the class H. An alternative mapping could be "H EQ public health animal health", but only if subordinate classes within class H apply to animal health as well as human health.

Table 5 (continued)

Class	Found in the thesaurus	Mapping options
E100	Preferred term “institutions” with NTs “educational institutions”, “hospitals” and “clinics”.	“E100 EQ educational institutions” The marker “=” or “~” may be applied, depending on whether the thesaurus term is intended to be used in the same sense as the class E100.
H100	“institutions” with NTs “hospitals” and “clinics” but no concepts corresponding to health centres or medical institutions in general, nor for more specific institutions such as dental practices or fitness centres.	“H100 EQ hospitals clinics” <i>or</i> “H100 NM hospitals” <i>as well as</i> “H100 NM clinics” The marker “~” would be redundant, since all compound equivalence mappings are to some degree inexact. The class H100 is far from exactly equivalent to the combination of hospitals and clinics, because these two do not add up to the whole scope of health institutions. The second option, in which each of them is clearly marked as a narrower concept, conveys the mapping situation more clearly than the first.
E210	“training” with NTs “first aid training” and “fire safety training”, and RTs “education” and “continuing professional development”.	“E210 ~EQ training” <i>or</i> “E210 BM training” The decision hinges on whether E210 is intended to cover only training for jobs within the education sector (in which case the thesaurus concept “training” is a broader concept) or whether it covers all kinds of training, in which case the inexact equivalence relationship is more appropriate.
H210	“training” with NTs “first aid training” and “fire safety training”, and RTs “education” and “continuing professional development”. Also present is preferred term “medical education”.	“H210 NM first aid training” “H210 BM training” “H210 RM continuing professional development” “H210 RM medical education” In this case, it is plain that the thesaurus concept “training” is much broader than H210, which is clearly limited to the Health sector. However, nothing approximately equivalent to H210 has been found, and so the only question remaining is how thorough to be in establishing any or all of the possible mapping statements enumerated here.
E211	Preferred terms “teachers” and “training”.	“E211 EQ teachers + training” This is a somewhat ambiguous mapping, since the same combination could be used for training <i>by</i> teachers as well as for training <i>of</i> teachers. In the absence of a thesaurus concept “teacher training”, however, it may be the best option available.
H220	Preferred terms “medicine” and “vaccinations”.	“H220 NM vaccinations” <i>or</i> “H220 RM vaccinations” “H220 BM medicine” The narrower mapping to “vaccinations” looks safe, because a vaccination can always be seen as a form of preventive medicine. But if the thesaurus term has been used for veterinary vaccinations, lying outside the scope of class H in the classification scheme, the associative mapping would be more appropriate. The proposed broader mapping to “medicine” also needs careful checking, as this thesaurus concept might apply only to the curriculum subject of medicine and not to routine practice.
H230	No thesaurus concept remotely close to operations (in the medical sense) is found.	Usually it is better to have no mapping at all than to establish a misleading one.

13.2.3 Mappings for conversion of search expressions when the thesaurus is the target vocabulary

Recommendations a to c, as well as the examples in 13.2.2, apply also to mappings for conversion of search expressions.

When search expressions are undergoing conversion, compound as well as simple equivalence mappings may be applied automatically with little impairment of retrieval quality. Where hierarchical and associative mappings are also present, human mediation (e.g. selection between the alternative mappings available) is advisable. See also 12.2.

13.2.4 Mappings for conversion of indexing/cataloguing entries when the thesaurus is the source vocabulary

The usual aim of classification is to assign a single class code to a document, combining the main strands of its subject content in a way that describes their treatment in the document. This should not be done on the sole basis of index terms allocated for post-coordinate use, because they could be coordinated with the wrong syntax. For example, if the index terms assigned to a document include “women” and “executives”, the document could deal with the treatment of women by executives, or with the performance of women in executive roles, among other possibilities, and each of these possibilities would demand a different class code. Hence generation of a notation, derived from the co-occurrence of index terms alone, is not recommended.

13.2.5 Mappings for conversion of search expressions when the thesaurus is the source vocabulary

Table 6 provides examples of mapping from a thesaurus to the classification scheme featured in Figure 8 (which shows fragments rather than the whole of the scheme). The following recommendations apply.

- a) At least one mapping should be sought for each concept in the thesaurus.
- b) Additional mappings should be established wherever the concept is found in a pre-coordinated class that is enumerated in the monohierarchical scheme.
- c) When an equivalence mapping has been established, it is not usually necessary to include mappings to any subclasses of the class in question. Thus in example a in Table 6, if the mapping “education EQ E” is selected, it is unnecessary to establish a direct mapping from “education” to “E100” or any other class in the Education schedule. (This recommendation assumes that the search system is able to exploit the structure of the classification scheme to extend a search for a particular class to any of its subclasses.)
- d) In addition to the mappings established by inspection of the classes enumerated in the scheme, provision should be made for classes and notations synthesized according to the rules of the scheme. One type of provision is the mapping shown in example f in Table 6, where a specific mapping can be synthesized to match a thesaurus concept. Another is the extension of any search to identify appropriate strings embedded in synthesized notations. For example, the search for “medical training” might be extended to detect “H210”, not just as a whole notation but as a fragment of a longer notation such as “S210(H210)”. To maximize recall, searches should be extended in this way, even where an exact simple equivalence mapping can be established, as in examples a, d, e and g. However, expansion of searches to retrieve such fragments from synthetic notation strings should be done with extreme care, as the rules for synthesis and parsing are often much more complex than this example implies.
- e) Where a query contains more than one thesaurus concept, unless the originator is able to check the outcome these should be converted separately rather than seeking to combine them in one synthesized notation. For example, if the query includes the terms “teachers” and “training”, conversion to “E211” would be inadequate, because it is not clear whether the user is seeking information about the training of teachers, or training *by* teachers.
- f) To maximize precision and recall, it is advisable to allow human mediation (or some other form of checking) in the search process (see 12.2).

Table 6 — Sample mappings to a classification scheme from a thesaurus

	Thesaurus concept	Found in classification scheme	Mapping options and discussion
a	education	E Education	"education BM E" or "education EQ E" The choice depends on whether the thesaurus concept is intended to be used in the same sense as the class E. This can be established by closer examination of both vocabularies and preferably also of the way they have been used in classifying/indexing the collections to be searched. Since the class E is very broad, encompassing institutions as well as the process of education, the hierarchical mapping is likely to be more appropriate. But if equivalence is chosen, optionally the marker "~" or "=" may be applied.
b	defence sector	D Defence	"defence sector BM D" or "defence sector EQ D" The choice depends on whether the class D is intended to be used in the same sense as the thesaurus concept. Neither of these mappings would be appropriate if D has a subclass that matches the scope of "defence sector" exactly.
c	institutions	E100 institutions H100 institutions D100 institutions	"institutions EQ E100 H100 D100" If the scheme has in it additional hierarchies for other sectors such as Transport, Sport, etc., and each of these has a class for institutions, these classes too should be included in the mapping. An alternative approach is to establish separate hierarchical mappings to all the relevant classes, for example "institutions NM E100", "institutions NM H100", etc.
d	educational institutions	E Education E100 institutions	"educational institutions EQ E100" Optionally the marker "~" or "=" may be applied, after inspection of the scope notes, narrower terms and other contextual information in both vocabularies.
e	schools	E110 schools	"schools EQ E110" Optionally the marker "~" or "=" may be applied, after inspection of the scope notes, narrower terms and other contextual information in both vocabularies.
f	diving schools	E110 schools S Sports S100 institutions S110 training centres S200 activities S245 swimming and diving	"diving schools BM S110:245" This mapping takes advantage of a rule within this classification scheme that allows a notation for institutions specific to a particular sport to be synthesized in the manner shown. The appropriate mapping type in this case is hierarchical, since the synthesized notation includes training centres for swimming as well as diving. A mapping to E110 might be considered, but rejected because the rules of the scheme make it clear that diving schools do not come within this class.
g	medical training	H Health H210 training	"medical training EQ H210" This mapping is likely to apply provided the scheme does not also include a pre-coordinated class incorporating some aspect of medical training. For example, in the Defence hierarchy, there could be a class such as D210(H210) enumerated for the specialist medical training given to combat troops. The above mapping would in this case be modified to read: "medical training EQ H210 D210(H210)"
h	training	E210 training E211 teacher training H210 training D210 training S210 training	"training EQ E210 H210 D210 S210" or all of the following: "training NM E210", "training NM H210"; "training NM D210", "training NM S210" It is not necessary to include any direct mapping to E211, assuming that E211 is already included implicitly in the superordinate class E210.

13.3 Mapping more complex classes

Classification schemes, especially large ones, often include classes more complex than the examples shown in 13.2. For these, an appropriate mapping statement can sometimes require a combination of the symbols + and |.

EXAMPLES

	Found in classification scheme	Found in thesaurus	Mapping statement
a	T563 inland waterway transport	rivers canals transport	T563 EQ (rivers canals) + transport
b	629.276, with caption “safety accessories”, within the hierarchy of “motor land vehicles, cycles”. The class also has a note: “Including air bags, bumpers, mirrors, seat belts, windshield wipers and washers”.	windscreen wipers air bags seat belts restraints (vehicle safety) motor vehicles cycles	629.276 EQ (motor vehicles cycles) + (windscreen wipers air bags seat belts restraints (vehicle safety))

Example b illustrates the need to check superordinate and subordinate classes as well as accompanying notes, when determining the scope of a particular class. It also raises the question of how reliably such a complex mapping statement would be interpreted in real retrieval situations. For example, Boolean operators are not usually admissible or meaningful when applied to index terms in metadata. And not all search engines will reliably accept nested parentheses and/or distinguish between those marking term qualifiers and those delimiting the search syntax.

The definition of some classes can involve exclusion of certain aspects. For example, the DDC class 331.8811 covers “labor unions in industries and occupations other than extractive, manufacturing, construction”. The question arises of whether to map this class to a combination such as “trades unions NOT (mining OR manufacturing OR construction)”. A mapping like this would have very limited usefulness. If used to convert cataloguing metadata to index terms from a thesaurus, a thesaurus term such as “labor unions” or “trades unions” could certainly be chosen, if present in the thesaurus. But current metadata standards provide no way of representing the exclusion of the other concepts.

If applied to a search statement, the negation could result in exclusion of some relevant documents, such as an article discussing the construction of premises for trades union activities. Like any use of the Boolean NOT in searching, human mediation is generally necessary to evaluate and adjust the search strategy.

The following recommendations apply to complex mappings.

- When a mapping statement incorporates a combination of different symbols, parentheses should be used to clarify the intended syntax.
- Complex mappings typically give inexact results, and so are most useful in applications with human mediation to separate the irrelevant from the relevant results.
- Mappings that involve negation or exclusion of certain concepts are not recommended except in circumstances where search results can be evaluated and the search strategy adjusted.

14 Techniques for identifying candidate mappings

14.1 General procedure

Traditionally the identification of mappings is an intellectual process. It needs one or more experts familiar with the relevant subject field(s), fluent in the language(s) of the vocabularies to be mapped, and having a good understanding of the structure and conventions of the vocabularies. The procedure is as follows.

- a) If more than two vocabularies are involved, the most appropriate structural model (see Clause 6) should be determined. After that is done, the work is usually allocated so that each expert is mapping between only one pair of vocabularies at any one time.
- b) The direction of the mappings should be chosen, as well as the range of mapping types to be employed. If mappings are needed in both directions, it may be found convenient to prepare all the mappings systematically in one direction first, and then to undertake all the mappings in the opposite direction.
- c) To encourage efficiency as well as thoroughness, the source vocabulary should be available to the expert in a sequence that supports systematic work, for example, in hierarchies organized by discipline or domain. The sequence of disciplines may be organized according to the availability of resources and the priorities of the application.
- d) The expert should work systematically concept by concept, looking up the target vocabulary to establish the appropriate mapping(s).
- e) In the most basic of systems it may be necessary for the expert to key in each mapping statement manually. If this is done, each term should be automatically validated against the source and target thesauri, as appropriate, to ensure it is a correctly spelt preferred term. (If one of the vocabularies is a classification scheme, validation applies to the notation rather than a term or caption. In the case of a subject headings scheme, the headings should be validated. Where synthesized notations or newly pre-coordinated headings are admissible, effective validation requires an algorithm.) In more sophisticated systems, data entry can be automated or partly automated; for example, by populating a template with all the preferred terms from the source thesaurus in a systematic sequence and/or by enabling completion of mapping statements by clicking on selected terms in a display of the target thesaurus. With all forms of automation, it is important to enable the expert to over-ride wrong term selections that were automatically entered.
- f) If mappings in the opposite direction are needed, the roles of source and target vocabulary should be reversed, and the same procedure repeated. The existing mappings (from Vocabulary A to Vocabulary B) may be exploited when mapping from B to A, in the following way.
 - Wherever exact equivalence has been reliably identified, the same mapping may be entered automatically in the opposite direction.

EXAMPLE 1

If “antennas” in Vocabulary A has an exact equivalence (see 11.2) mapping to “antennae” in Vocabulary B, the reverse mapping “antennae =EQ antennas” can be accepted without checking.

- Where inexact equivalence has been established, the same mapping may be offered to the expert with its direction reversed for checking and approval or rejection/modification.

EXAMPLE 2

If “repossession” in Vocabulary A has an inexact equivalence (see 11.3) mapping to “foreclosures” in Vocabulary B, the reverse mapping “foreclosures ~EQ repossession” could be offered to the expert for checking and approval or rejection/modification.

- Where intersecting compound equivalence (see 8.3.2) has been established, two or more separate hierarchical mappings from broader to narrower may be offered to the expert with direction reversed for checking and approval or rejection/modification.

EXAMPLE 3

If “prize-winning swimmers” in Vocabulary A has been mapped to “prize-winners + swimmers” in Vocabulary B, the reverse mappings “prize-winners NM prize-winning swimmers” and “swimmers NM prize-winning swimmers” could be offered to the expert for checking and approval or rejection/modification.

- Where cumulative compound equivalence (see 8.3.3) has been established, two or more separate hierarchical mappings from narrower to broader may be offered to the expert with direction reversed for checking and approval or rejection/modification.

EXAMPLE 4

If “bikes” in Vocabulary A has been mapped to “motorbikes | pedal cycles” in Vocabulary B, the reverse mappings “motorbikes BM bikes” and “pedal cycles BM bikes” could be offered to the expert for checking and approval or rejection/modification.

- Where a hierarchical mapping (see Clause 9) from a narrower to a broader concept has been reliably established, a hierarchical mapping from the broader to the narrower concept may be offered to the expert for checking and approval or rejection/modification.

EXAMPLE 5

If “pebbles” in Vocabulary A has a hierarchical (narrower to broader) mapping to “stones” in Vocabulary B, the reverse mapping “stones NM pebbles” could be offered to the expert for checking and approval or rejection/modification.

- Where a hierarchical mapping (see Clause 9) from a broader to a narrower concept has been reliably established, a hierarchical mapping from the narrower to the broader concept may be offered to the expert for checking and approval or rejection/modification.

EXAMPLE 6

If “cattle” in Vocabulary A has a hierarchical (broader to narrower) mapping to “cows” in Vocabulary B, the reverse mapping “cows BM cattle” could be offered to the expert for checking and approval or rejection/modification.

- Wherever an associative mapping (see Clause 10) has been established, the same mapping may be offered to the expert for checking and approval or rejection/modification.

EXAMPLE 7

If “promotion” in Vocabulary A has an associative mapping to “publicity” in Vocabulary B, the reverse mapping “publicity RM promotion” could be offered to the expert for checking and approval or rejection/modification.

14.2 Computer assisted direct matching

It is sometimes possible to automate the process described in 14.1 d), at least in part, using a matching algorithm.

If both of the vocabularies are thesauri with at least one language in common, the general procedure is to compare each term in the source thesaurus with all the terms in the same language in the target vocabulary. Where a match is found, this may be accepted as a candidate mapping. Where a concept in the source vocabulary has non-preferred terms, a different candidate mapping could potentially be found for each of these, as well as for the preferred term.

Additional candidate mappings may be found by stripping any qualifiers off the terms and repeating the process.

If the thesaurus has more than one language in common, the process should be repeated for each one of these.

If a candidate mapping has not yet been found for a concept by any of the above means, the terms in both target and source languages may be stemmed and the process repeated for each language. Additional natural language processing techniques may be applied, but a description is outside the scope of this part of ISO 25964.

The candidate mappings identified by the matching processes described should be assembled for review by an expert. For each concept in the source vocabulary, the expert should be able to view the complete record (including scope note, broader and narrower terms). The candidate mappings for the concept should be presented in ranked order, to suit the convenience of the expert in making comparisons and finding exact equivalents first. For example, matches on preferred terms are usually ranked highest, matches on non-preferred terms lower, followed by matches after stripping off qualifiers, and finally the matches after stemming. The viewing interface should make it easy to check the complete context of each concept identified in the target vocabulary. It should also support the expert in selecting the appropriate type of mapping for the candidate(s) he approves.

Only after an expert has reviewed them should any of the candidate mappings be approved and established.

If one of the vocabularies to be mapped is not a thesaurus, the matching procedure should be adapted. The captions in a classification scheme or category labels in a taxonomy are much less reliable than thesaurus terms, for unambiguous description of the class or category (see 17.2.2). In the case of a subject heading scheme, headings that consist of a single term may be treated like thesaurus terms, but compound headings need special treatment (see 20.3).

The matching procedures outlined may often be extended or adapted to suit the context of a particular vocabulary or application.

14.3 Co-occurrence mapping

This method exploits the co-occurrence of terms from different schemes in the same metadata or catalogue record. It relies on having at least one large existing collection already indexed with both vocabularies. Taking Vocabulary A as source, the metadata of all records indexed with a particular term are analysed to discover which terms from Vocabulary B co-occur. The highest ranking of these may be used as candidates from which to select valid mappings. Only after an expert has reviewed them should any of the candidate mappings be approved and established. Description of the co-occurrence algorithm is outside the scope of this part of ISO 25964.

14.4 Other methods

This is a fast developing field and other methods may well emerge. This part of ISO 25964 does not exclude any emerging technologies, but recommends that for reliable mappings of a good quality, any candidate mappings that are generated automatically should be subject to review by an expert.

15 Managing the data

15.1 Types of data to be recorded

15.1.1 Overview

Mappings data can be recorded at three levels:

- a) data describing the mapping of individual concepts;
- b) data describing the mapping of groups of concepts between all or part of two controlled vocabularies (for example, a complete set of mappings between two thesauri, or the set of mappings from one microthesaurus to a subject heading scheme, or the mappings from one specific concept group in a thesaurus to any other vocabulary);

NOTE Exceptionally, a mapping set can involve a combination of two or more **target vocabularies**, where one or more of them act(s) as qualifier to the other (see 8.3.4).

- c) data describing a mapping cluster, that is to say a coordinated set of mappings between more than two controlled vocabularies (for example, the set of mappings between one hub vocabulary, as illustrated in Figure 2 of 6.4, and three other vocabularies).

15.1.2 For mappings between individual concepts

At this level, data should specify:

- a) the two or more concept identifiers, preferred terms or notations to be mapped (just two in the case of a simple mapping, but more than two where compound equivalence (see 8.3) applies);
- b) the type of mapping (see Clauses 7 to 10) between the concepts, including, if applicable, compound operators ('|' or '+') combining the target concepts (see 8.3) and an exact or inexact marker (see Clause 11);
- c) the validity period of the mapping, when not implicit in the specification of the vocabularies being mapped; and

NOTE This is required when concepts in a particular controlled vocabulary are added, deleted or changed in scope and should take account of any history notes attached to individual concepts in the vocabularies concerned.

- d) any annotations with additional optional information (e.g. to control ranking or to note level of confidence in the quality of the mapping).

NOTE A full enumeration of such annotations is application specific and is beyond the scope of this part of ISO 25964.

Mapping between individual concepts is meaningful only in the context of the controlled vocabularies in which the scopes of these concepts are explicitly or implicitly defined. These vocabularies should therefore be identified, as discussed in 15.1.3.

For each mapping, the characteristics described in 15.1.3 b) to d) should also be specified. However, it may be convenient to specify them at the group level rather than individually. For some mappings, it may be appropriate also to specify on an individual basis the characteristics described in 15.1.3 e) to j).

15.1.3 For a set of mappings between groups of concepts

At the level of mappings between groups of concepts forming all or part of two controlled vocabularies, data should specify:

- a) identification of the set of mappings, including, where appropriate, identification of the version or the date when the mappings were last modified;
- b) the source vocabulary (including its date and/or version);
- c) the target vocabulary (including its date and/or version);
- d) the direction of the mappings, if one-way;
- e) any restrictions on the scope of the mappings, for example, if the complete source and target vocabularies were not included, then the parts used or the principle of selection should be specified; if the mappings were designed for a particular purpose or application, that should be specified;
- f) the process by which the mappings have been created, e.g. whether by human intellectual analysis, automated matching of terms, co-occurrence in a collection of resources indexed by both vocabularies, or some combination of such techniques;
- g) identification of the person(s) or organization responsible for maintaining and/or publishing the mappings;
- h) how the mappings can be obtained or accessed (see also 15.3.4);
- i) any related mappings or cluster of mappings to which this mapping set belongs, as discussed in 15.1.4; and
- j) any copyright, confidentiality or security restrictions on use or exchange of the mappings.

Optionally, notes concerning the level of confidence in the quality of the set of mappings may be added.

15.1.4 For mapping clusters

When a mapping forms part of a managed cluster of mappings between more than two controlled vocabularies, the following data about the cluster should be specified:

- a) identification of the cluster;
- b) the structural model of the cluster (see Clause 6) and, in the case of a hub structure, identification of the controlled vocabulary used as hub;
- c) the intended purpose or application of the cluster;
- d) identification of the person(s) or organization responsible for maintaining and/or publishing the cluster; and
- e) any copyright, confidentiality or security restrictions on use or exchange of the cluster.

Optionally, notes concerning the level of confidence in the quality of the mapping cluster may be added.

15.2 Storing the data

15.2.1 Organizational aspects

A mapping may be stored:

- within the system managing the source controlled vocabulary;
- within the system managing the target controlled vocabulary;
- in a system dedicated to managing the mappings, but independent of any system managing either the target or the source controlled vocabularies; or
- in a system managing all the source and target controlled vocabularies, as well as the mappings between them.

Commonly the choice of organizational mode is determined by the data ownership and business model, rather than by technical criteria. However, some implications of this choice are discussed in 15.3.

15.2.2 Technical aspects

Mappings, sets of mappings and clusters are typically stored in a database. For each mapping the respective characteristics detailed in 15.1 should be specified. Each individual mapping should be detailed as a relationship or as a rule.

Storage of these mappings may be done using a database technology, such as a relational database, an XML database, a rule database or an RDF store. Each of these methods requires a dedicated schema, satisfying the following requirements:

- the schema should be able to specify the different mapping types described in Clauses 7 to 10, including compound equivalence (see 8.3) if used;
- the schema should be able to specify annotations on stored mappings;
- the schema should be able to specify the characteristics of any mapping sets (see 15.1.3); and
- the schema should be able to specify the characteristics of any mapping clusters (see 15.1.4).

As yet there is no standard schema that fully complies with this part of ISO 25964, and the development of such a schema is not within scope. However, this is a rapidly evolving field and implementers of this part of ISO 25964 should be alert to developments among interested parties, e.g. the SKOS^[37] user community.

The schemas developed for storage purposes may also be used or adapted to enable publication of the mappings. A SKOS-compliant format^[37] is recommended if use in the Semantic Web is desired.

15.3 Maintaining the mappings data

15.3.1 General

Maintenance of mappings data is particularly complex. Care should be taken to use tools and procedures that are sustainable for maintaining the mappings. In all cases, the mapping tools need to have access to the source and target vocabularies.

During the maintenance of a controlled vocabulary, if the vocabulary is involved in a set or cluster of mappings, the scope of the concepts involved in mappings should not be allowed to change, without also amending the corresponding mappings. ISO 25964-1:2011, 13.6.4, shows the types of change that can occur.

15.3.2 Changes in the source or target vocabularies

When either the source or the target vocabulary is modified, the mappings to or from its concepts should be checked. A list of differences between the previous version and the new version of the vocabulary, as recommended in ISO 25964-1:2011, 13.6.5.2 and 13.6.5.3, can help the checking process, especially if the changes that affect the scope of the mapped concepts have been identified. Automatic tests can be unreliable as they do not usually indicate, for example, whether a new non-preferred term is simply a spelling variant or represents an extension of the scope of the corresponding concept.

After all the mappings have been corrected, a new version should be issued of any set or group of mappings to which the mappings belong.

Table 7 shows the main types of action required during or after the updating of a source vocabulary, while Table 8 applies during or after the updating of a target vocabulary.

Table 7 — Implications for mappings of changes in the source vocabulary

Change in the source vocabulary	Implications for mappings
New concept is introduced	A new mapping is required, assuming there is a corresponding concept in the target vocabulary.
Concept is removed	The mapping of the source concept should either: — get a validity annotation to indicate its limited use; or — be removed in a new version of the mapping set or cluster.
Complex replacement of a concept (e.g. a split or merger)	All relevant mappings should be checked and corrected as necessary. They should not be automatically updated.
Changes affecting the scope of an existing concept	Existing mappings from that concept should be checked and corrected as necessary.

Table 8 — Implications for mappings of changes in the target vocabulary

Change in the target vocabulary	Implications for mappings
New concept is introduced	Existing mappings remain valid but are likely to need improvement. Any mappings to the broader, narrower or associatively related concepts of the new concept should be checked and corrected where appropriate. Additionally, the source vocabulary should be searched for any concepts that could usefully be mapped to the new target concept.
Concept is removed	Any simple mappings to the removed target concept should be removed and, if necessary, replaced by more appropriate mappings from the source concept(s) involved. Any compound mappings to the removed target concept should be evaluated and corrected or replaced as necessary.
Complex replacement of a concept (e.g. a split or merger)	All relevant existing mappings should be checked and corrected or replaced with new ones as necessary. They should not be automatically updated.
Changes affecting the scope of an existing concept	Existing mappings to that concept should be checked and corrected as necessary.

15.3.3 Other mapping changes

Even if the source and target vocabularies are unchanged, a need to review and adjust the mappings can arise. If the mappings are revised in any way, the changes should be communicated to users and reflected in any information retrieval systems that are dependent on the mappings. It should be remembered that an individual mapping may belong to more than one set or group of mappings, and all sets affected by the changes should be updated accordingly.

15.3.4 The impact of mapping changes

Since mappings and sets of mappings are developed primarily to support information retrieval, as described in Clause 5, any changes in them need to be reflected in the indexing and search applications that exploit them. (See also the explanations and examples in Clause 12.) A list of the changes affecting each group or set of mappings should be issued to enable:

- a) reconfiguration of any indexing engine that uses the mappings;
- b) an assessment of the impact on the metadata of documents already indexed via the mappings; and
- c) reconfiguration of any search systems using the mappings. Where the mappings apply only to search terms (and not to index terms), a negative impact on retrieval performance is unlikely.

If the mappings are published externally, the list of changes should be published also.

Typically, significant time is needed before all indexed content has been updated with the newest versions of updated vocabularies and mapping sets. Therefore, older versions of vocabularies and mapping sets should remain available until such time as they are no longer used [sometimes the mappings need annotations that constrain them in time or to particular versions; see 15.1.2 d)].

16 Display of mapped vocabularies

16.1 General

No one type of display meets all the likely needs. New styles of display for retrieval applications are continuously evolving. Apart from the general principles outlined here, this part of ISO 25964 does not seek to constrain the presentation of mappings data to end users. The displays described in 16.2 to 16.4 are intended mostly for managing the development and maintenance of mappings.

For some purposes it is sufficient to display only one of the mapped vocabularies (without revealing the mappings) and, if this is a thesaurus, the display should be as described in ISO 25964-1:2011, Clause 12.

End users should not be expected to recognize and correctly interpret the tags described in Clause 4. These are intended for use by trained personnel.

For some applications, the mappings are used mainly by computers and display of them could be confusing as well as unnecessary. For example, humans need not see the mappings at all if the terms they enter into a search system are converted automatically into the equivalents in another language. Even if human mediation is invoked (see 12.2), the choices may be presented to the user in a way that does not require him to be aware of the mapping process. For example, mapped terms from the target vocabulary may be presented in the style of a tag cloud (comparable to the style shown in some social networking applications) without explicit designation of the type of mapping in each case.

However, a display of mappings can be useful to enable browsing of the vocabularies available. And some form of display may be appropriate when users need to choose between alternative mappings.

In applications where user selection of search keys (such as terms of notations) is encouraged, support via direct search, hyperlinked navigation, user-friendly hierarchical displays and other visual clues should be provided as much as resources permit.

When a large number of mappings are presented for user selection, they should be clustered in helpful groups to avoid confusion. Depending on the application and context, clustering may be by type of mapping, by degree of equivalence and/or degree of overlap (in the case of inexact equivalence), by target vocabulary, by target entity type (in the case of name authority lists) or by some other attribute that makes sense to users.

When the source or target vocabularies are presented in more than one language, the browsing interface should allow the user to switch to the language of his choice, and preferably all the relevant terms should be available in that language.

A display of mappings should also be available to the persons who establish and maintain them. The displays shown in 16.2 and 16.3 may be found useful in this context.

16.2 Single record display

A single record display is often required for users trained in the use of mappings, especially during the process of establishing and maintaining the mappings. This type of display assumes one vocabulary as the source vocabulary, from which mappings to the target vocabulary/ies have been established or are in the process of being established.

The basics of the display should be as described in ISO 25964-1:2011, 12.2, augmented by mappings using the conventions outlined in Clause 4 of this part of ISO 25964, and explained in more detail in Clauses 7 to 13. Optional elements of the source vocabulary (such as Top Term and Definition) may be included if required.

The style of display depends on how much the different types of mapping and degrees of equivalence have been differentiated. To illustrate the options, the displays in Tables 10 and 11 will draw on the data in the records excerpted from three thesauri as shown in Table 9. Mappings are in the course of establishment from one source thesaurus to two target thesauri. In the first target thesaurus, there is an exactly equivalent concept, but in the other one the only candidates are either much broader or much narrower, possibly suitable for hierarchical mappings.

In this situation, mapping can be done with more or less differentiation as illustrated in Table 10. VOC1 and VOC2 are the identifiers of the two target thesauri. The first column (undifferentiated mapping) shows a minimal approach, in which only one mapping per target vocabulary has been identified for each concept from the source vocabulary, and the type of mapping has not been specified. In the second column (differentiated mapping), more trouble has been taken to identify and display all the mappings that might be useful from a particular concept. The latter approach supports discrimination between alternative mappings when an exact equivalent concept is not available. This style of display should also accommodate any compound equivalence mappings that apply.

**Table 9 — Corresponding records from three thesauri
for which mappings are under construction**

Source vocabulary	Target vocabulary 1	Target vocabulary 2
dairy products	milk products	animal products
UF <i>milk products</i>	BT animal products	BT products
BT animal products	food products	NT leather
NT butter	NT cheeses	meat
cheese	ice cream	milk
cream	RT dairies	RT livestock industry
milk		
RT dairies		

**Table 10 — Single record display for two different styles of mapping
(from one source thesaurus to two target thesauri in both cases)**

Undifferentiated mapping	Differentiated mapping
dairy products	dairy products
UF <i>milk products</i>	UF <i>milk products</i>
BT animal products	BT animal products
NT butter	NT butter
cheese	cheese
cream	cream
milk	milk
RT dairies	RT dairies
VOC1 milk products	VOC1 =EQ milk products
VOC2 animal products	VOC2 BM animal products
	VOC2 NM milk

Table 11 extends the display even further by including additional valid mappings to concepts in VOC1. The hierarchical/associative mappings to “animal products”, “food products”, “cheeses”, “ice cream” and “dairies” are not wrong, but they are generally considered unnecessary to establish or display because they can be derived automatically from the structure of internal relationships within VOC1, provided that an exact equivalence relationship has been established to “milk products”.

**Table 11 — Single record display with differentiated mapping,
extended to include redundant mappings**

dairy products
UF <i>milk products</i>
BT animal products
NT butter
cheese
milk
RT dairies
VOC1 =EQ milk products
VOC1 BM animal products
food products
VOC1 NM cheeses
ice cream
VOC1 RM dairies
VOC2 BM animal products
VOC2 NM milk

The differentiated style without redundancy, as illustrated in the second column of Table 10, is recommended as the most powerful and flexible of the options.

The undifferentiated style may be acceptable as a low-cost alternative in applications that do not require high quality mappings.

As illustrated in Tables 10 and 11, the tag or field label for each mapping in a monolingual context has up to three components: firstly the identifier for the target thesaurus, secondly a tag for mapping type and optionally a tilde or equals sign to show inexact or exact equivalence where applicable. All the tags and symbols are set out in Clause 4 of this part of ISO 25964 plus ISO 25964-1:2011, Clause 3. The identifiers of the source and target thesauri should be chosen to avoid confusion with any of the tags.

If the target thesaurus is multilingual, the identifier for it may additionally include an indicator of the language version within it. In the example, "VOC1" should become "VOC1en" and "VOC1 EQ" should become "VOC1en EQ".

If mappings to another type of vocabulary are included (instead of or as well as VOC1 and VOC2 in Tables 9 to 11), the following target components should be shown in each mapping:

- for a subject heading scheme, the appropriate heading(s);
- for a name authority list, the appropriate preferred name(s) or identifier(s);
- for a classification scheme, the appropriate notation(s) and corresponding caption(s);
- for a taxonomy, the appropriate category label(s) and if this is not unique in the taxonomy, some means of disambiguation.

If the mappings are intended for use in more than one direction, a comparable display should be available for each vocabulary that will be used as source.

16.3 Complete displays based on one of the vocabularies

16.3.1 Alphabetical displays

If the source vocabulary chosen for the alphabetical display is a thesaurus, the single records within a complete display should be as described in 16.2. If all of the other vocabularies are monolingual thesauri, the complete display will then be as shown in Figure 9. This figure is based on the same three thesauri as appear in Table 9. A comparable display should also be available for each thesaurus that will be used as source vocabulary, i.e. VOC1 and VOC2 in this example.

If the source vocabulary is a name authority list, an alphabetical display may be prepared in a similar way.

agrochemicals		cheese	
UF	<i>agricultural chemicals</i>	BT	dairy products
NT	fertilizers	VOC1 =EQ	cheeses
	pesticides	VOC2 BM	animal products
VOC1 =EQ	agrochemical products	VOC2 RM	milk
VOC2 =EQ	agricultural chemicals		
animal products		chickens	
NT	dairy products	UF	<i>hens</i>
	eggs	BT	poultry
	leather	VOC1 =EQ	chickens
	meat	VOC2 =EQ	chickens
	wool		
VOC1 =EQ	animal products		
VOC2 =EQ	animal products		
butter		cream	
BT	dairy products	BT	dairy products
VOC1 BM	milk products	RT	milk
VOC2 BM	animal products	VOC1 BM	milk products
VOC2 RM	milk	VOC2 BM	animal products
		VOC2 RM	milk
cattle		dairy products	
BT	livestock	BT	animal products
RT	milk	NT	butter
VOC1 =EQ	cattle		cheese
VOC2 ~EQ	cows		cream
			milk
cereals			dairies
BT	plant products	RT	dairies
VOC1 =EQ	cereals	VOC1 =EQ	milk products
VOC2 BM	plant products	VOC2 BM	animal products
		VOC2 NM	milk

Figure 9 — Alphabetical display with mappings to two target vocabularies

16.3.2 Systematic displays

Systematic displays with mappings may be useful if the source thesaurus habitually uses a classified display. They are similarly applicable if the source vocabulary is a classification scheme (including schemes used for records management), taxonomy or subject heading scheme, since these are generally presented in systematic rather than alphabetical order.

16.4 Crosswalks

For many applications the display of internal relationships within the source or target vocabularies is unnecessary. A correspondence table or crosswalk, showing only the cross-vocabulary mappings, may be more appropriate. Table 12 shows mappings from one source thesaurus to two target thesauri, based on the same data as in Table 9 and Figure 9.

Table 12 — Crosswalk with mappings to two target vocabularies

Source thesaurus concept	Mappings from source to Voc 1		Mappings from source to Voc 2	
	Mapping type	Concept	Mapping type	Concept
agrochemicals	=EQ	agrochemical products	=EQ	agricultural chemicals
animal products	=EQ	animal products	=EQ	animal products
butter	BM	milk products	BM RM	animal products milk
cattle	=EQ	cattle	~EQ	cows
cereals	=EQ	cereals	BM	plant products
cheese	=EQ	cheeses	BM RM	animal products milk
chickens	=EQ	chickens	=EQ	chickens
cream	BM	milk products	BM RM	animal products milk
dairy products	=EQ	milk products	BM NM	animal products milk

When interpreting the data in a table such as this, there is an important limitation to note. The mappings are in each case from the concept in column 1 (drawn from the source thesaurus) to the corresponding concepts in columns 3 or 5. Unless all of the mappings included are of the exact equivalence type, the table does not show mappings from one target thesaurus to another. In the penultimate row of the table, for example, we see the Voc 1 concept “milk products” in the same row as the Voc 2 concepts “animal products” and “milk”, but without more information the type of mapping that could apply between these concepts cannot be known precisely.

17 Classification schemes

17.1 Key characteristics and background

17.1.1 General description

A classification scheme is a system of classes or categories into which information resources of all kinds may be placed. As described in this clause, it can be used either for arranging items on the shelves of libraries or for arranging the surrogates of items in catalogues. The schemes used for records management, and for organizing web pages and portals, are described separately in Clauses 18 and 19 respectively.

The basic approach of classification is to organize concepts into classes. The classes are subdivided into narrower classes (see 17.2.3), and these into yet narrower classes down to the level of specificity required.

Each class has a caption (see 17.2.2) and sometimes a longer description to define the scope of the class. Functioning like a thesaurus scope note, the description often lists topics that are included and provides “See also” cross-references to other classes where related topics are located.

In a subject classification scheme (see 17.1.5.1) each class also has a notation (see 17.2.1), typically a short string of alphabetical and/or numerical characters and/or other symbols.

The classification approach typically provides for pre-coordination of concepts, and has been adopted by numerous other types of vocabulary, including many taxonomies (see Clause 19) and the schemes used in

records management systems (see Clause 18). Since pre-coordination presents problems for interoperability, the implications and examples for all such vocabularies are discussed further in Clause 13.

Some examples of different styles of classification scheme appear in Figures 10 and 11. In both cases only brief extracts are shown, and information other than the caption (see 17.2.2) and notation (see 17.2.1) has been excluded.

17.1.2 Scope and role in information retrieval

The traditional purposes of classification schemes are to organize information resources, whether in hard copy or electronic, and to allow retrieval by browsing the shelves or by browsing a classified catalogue.

Today's electronic networks make it possible to extend the use of classification schemes to search across multiple collections and databases.

17.1.3 Origin and development

Classification schemes have been used throughout the history of libraries. Many of the schemes in use today derive some structure and practices from schemes developed a century ago or more, especially Melvil Dewey's *Dewey Decimal Classification* (DDC)^[16], C.A. Cutter's *Expansive Classification*, the *Universal Decimal Classification* (UDC)^[32] by LaFontaine and Otlet, the *Library of Congress Classification* (LCC)^[21] by Hanson and Martel, and J. D. Brown's *Subject Classification* (SC)^[14]. These early schemes were mostly enumerative (see 17.1.5.2), although the UDC brought in some synthetic principles from the start. The idea of faceted classification came later, with Ranganathan's *Colon Classification* (CC)^[27] and H.E. Bliss's *Bibliographic Classification* (BC)^[11]. Since the 1950s, some of the techniques of faceted classification have been gradually introduced into most of the enumerative schemes.

17.1.4 Vocabulary control

Whereas a thesaurus exercises vocabulary control by designating a preferred term for each concept, a classification scheme gives a unique notation for each class. Being independent of the language(s) of the scheme, the notation serves as an unambiguous label to represent the scope of the class.

Where an alphabetical index (see 17.2.4) is available, this normally provides extra entry points to the classes from alternative terms and ways of expressing the corresponding concepts. The index enables natural language access to the notation codes that constitute the artificial controlled language of the scheme.

17.1.5 Types of classification schemes

17.1.5.1 General

The type of classification scheme used for records management is considered in Clause 18, and the type used for organizing web pages and portals in Clause 19. The present clause deals primarily with classification schemes used by libraries (whether physical or electronic) or bibliographic databases, in which the objective of the scheme is to enable access to documents according to their subject matter. Such schemes are sometimes called "subject classification schemes".

17.1.5.2 Enumerative versus synthetic schemes

All subject classification schemes set out a number of classes in a schedule or set of schedules. The process of setting them out is often known as enumeration. A scheme that explicitly lays out all the classes available to the user is known as an enumerative classification scheme.

When the scheme needs to enumerate a class combining two or more concepts that belong in different classes, a decision has to be made about where to place it. Usually it is located under one of the constituent classes, according to rules specific to the scheme in question.

As an alternative to detailed enumeration, some schemes include rules for combining classes so that the user can synthesize compound notations from the more elementary notations of the constituent classes. The scheme can therefore provide for very many more subjects than it is practical to enumerate. Schemes with this capability are known as synthetic or analytico-synthetic schemes.

If the notation is regarded as constituting the basic controlled vocabulary of a classification scheme, in an enumerative scheme this vocabulary is finite and explicitly laid out. For a synthetic scheme, the vocabulary is not so tightly bounded. However, the rules of synthesis limit the style and structure of the synthesized notations. Usually it is possible to decompose these into the constituent notations enumerated in the schedules (see also 13.2.1.)

Often classification schemes today incorporate a mix of enumerative and synthetic features.

17.1.5.3 Faceted classification schemes

Faceted classification schemes take the principles of analysis and synthesis one step further. Firstly subjects are analysed into simple concepts according to fundamental categories such as activities, entities, places, etc. These concepts are enumerated in the scheme as classes, each with an assigned notation. Then the notation for a complex subject is synthesized by combining the notation of the simpler concepts in accordance with rules for sequence, known as citation order. The sequencing rules are essential to ensure that all the documents on the same complex subject are given exactly the same notation and hence are brought together. Figure 10 illustrates a simple faceted classification scheme, in which the two facets employed are “organisms” and “processes”.

Classification schedule	
<i>(organisms)</i>	
A	mammals in general
AA	carnivores in general
AAA	leopards
AAB	lions
AAC	tigers
AB	herbivores in general
ABA	cattle
ABB	sheep
<i>(processes)</i>	
B	physiological processes in general
BB	digestion in general
<i>(organisms)</i>	
BB.AA	digestion in carnivores
BB.AAB	digestion in lions
BB.AB	digestion in herbivores
BB.ABA	digestion in cattle
BB.ABB	digestion in sheep
BC	respiration in general
<i>(organisms)</i>	
BC.AA	respiration in carnivores
BC.AAB	respiration in lions

Figure 10 — Example of a short, faceted classification with a simple notation

17.2 Semantic components and relationships compared with thesaurus components

17.2.1 Notation

17.2.1.1 Purposes

The prime functions of a notation are:

- a) to provide a systematic sequence that enables location of concepts within the scheme and documents classified by the scheme; and
- b) to serve as an unambiguous label for the intended concept.

17.2.1.2 General description

Notations are typically short strings of alphabetic or numeric characters or other symbols, or a mixture of these, that can be used to represent classes. Enumerative schemes lay out all the notation strings available to users. Synthetic schemes and faceted schemes provide rules that allow users to synthesize notation strings for concept combinations that have not been enumerated. Examples of notation are shown in Figure 10 and Figure 11. In Figure 11, the first nine classes are enumerated in the scheme; the last two were synthesized by following the UDC rules.

(084)	pictorial documents
(084.12)	photographs
5	mathematics and natural science
59	zoology
591.1	animal physiology
591.132	digestion
599	mammals
599.74	carnivores
599.742.71	"big cats"; lions; tigers
599.742.71:591.132	digestion in lions
599.742.71(084.12)	photographs of lions

**Figure 11 — Selected classes from a classification scheme (UDC)
with a notation that uses numbers and symbols**

Punctuation marks such as colons, commas, inverted commas, etc., are often used in notations, although practice varies widely. To interpret any such characters it is necessary to consult the rules of the scheme in question.

17.2.1.3 Implications for mapping classes to or from thesaurus concepts

Classes such as 591.132 (digestion) or 599 (mammals) in Figure 11 are relatively simple and so could readily find equivalent concepts, either exact or inexact, in a thesaurus with suitable scope (see Clauses 8 and 11). Depending on the relative specificity of the scheme and the thesaurus, some simple classes might alternatively need a hierarchical or associative mapping (see Clauses 9 and 10). A class such as 599.742.71(084.12), however, is so complex that it is unlikely to correspond to a single thesaurus concept. Such classes are very common in large classification schemes of all types, and to map them it is necessary to establish compound equivalences (see 8.3).

17.2.2 Captions

Figures 10 and 11 show a caption (sometimes called a class heading) alongside the notation for each class. The caption briefly conveys the scope of the class, but does not have to be unique because the user is expected to view it in the context of the superordinate class. For example, the classes “fungi” and “fruits” might both be divided into subclasses with captions “poisonous” and “edible”. The user is expected to understand that the concept behind “poisonous” is really poisonous fungi in the first case and poisonous fruits in the other. There is assumed to be no confusion in the location of classes on these topics, because they are given different notations. In this respect classification schemes differ from thesauri, in which the concepts are uniquely identified by terms rather than by notation (see ISO 25964-1:2011, 6.2.1.) The implications for mapping between a classification scheme and a thesaurus are:

- the caption alone is generally an insufficient indicator of the scope of a class; and
- mapping statements should be expressed using the notation to represent a class, rather than using its caption.

17.2.3 Hierarchies in classification schemes

The approach of dividing classes into narrower classes, and these into yet narrower classes, is fundamentally hierarchical. The hierarchical approach is often accentuated in the layout of a classification scheme, with the size and style of typeface varying as the level of subdivision proceeds and the use of indentation (as in Figure 10) to show which level has been reached. However, it is very important to note that the hierarchies of a classification scheme are not usually governed in the same way as hierarchical relationships in a thesaurus.

A thesaurus admits relationships as hierarchical only if they are generic, partitive or instantial (see ISO 25964-1:2011, 10.2). For example, since a chemical element is neither a type, nor a part, nor an instance of chemistry, a BT/NT relationship between these concepts is not applicable. However, in a library it is convenient to be able to find books about chemical elements in the Chemistry section. Therefore the subclass “chemical elements” may be admitted at the appropriate level within the hierarchy of the class “chemistry” in a classification scheme for use in libraries. The notation similarly shows chemical elements as belonging to chemistry.

It follows that when the classes of a classification scheme are mapped to the concepts of a thesaurus, even if every class finds an exact equivalence mapping, commonly the hierarchical structure between the classes is not the same as the hierarchical relationships between the thesaurus concepts. This should not be regarded as a problem. For purposes of information retrieval, a crosswalk between a thesaurus and a classification scheme needs mappings between concepts and classes, not between the respective internal relationships of the vocabularies.

The rules of citation order in a classification scheme are designed to lead to a single location for any compound class. This is convenient in physical libraries, where there needs to be one (and only one) place for every document, though documents that deal with more than one subject can have entries under more than one class in a classified catalogue. See ISO 25964-1:2011, 10.2.5 for a discussion of the ways in which thesauri can handle alternative hierarchical locations of concepts.

17.2.4 Alphabetical indexes to classification schemes

Most classification schemes have an index enabling access to the multiple locations at which a single concept can occur within the scheme. This is especially necessary for faceted classification schemes because, as illustrated in Figures 10 and 12, aspects of simple concepts, such as lions, are scattered when they appear in compound subjects but are not the first-cited elements.

Alphabetical index corresponding to the scheme in Figure 10	
carnivores : digestion : physiological processes	BB.AA
carnivores : mammals	AA
carnivores : respiration : physiological processes	BC.AA
cattle : digestion : physiological processes	BB.ABA
cattle : herbivores : mammals	ABA
digestion : physiological processes	BB
herbivores : digestion : physiological processes	BB.AB
herbivores : mammals	AB
leopards : carnivores : mammals	AAA
lions : carnivores : mammals	AAB
lions : digestion : physiological processes	BB.AAB
lions : respiration : physiological processes	BC.AAB
mammals	A
physiological processes	B
respiration : physiological processes	BC
sheep : digestion : physiological processes	BB.ABB
sheep : herbivores : mammals	ABB
tigers : carnivores : mammals	AAC

Figure 12 — Selected entries from a chain index

Whether the index is a chain index as in Figure 12 or uses the style recommended by ISO 999:1996, its function is only to provide access to the schedules of the scheme. Sometimes it does this by including entries for specific aspects of the scope of a class, that might not be explicitly mentioned in its caption or accompanying notes. For example, in the *DDC*, index entries leading to class 387.2 (Ships) include “Air-cushion vehicles — ocean”, “Boats”, “Merchant ships”, “Vessels (Nautical)”, etc. While establishing mappings between a classification scheme and a thesaurus, such index entries may prove helpful in clarifying the scope of the corresponding class. They may also be helpful in suggesting additional hierarchical mappings to and from thesaurus concepts. In any mapping statement, however, the class should be represented by its notation and not by any of its index entries.

17.3 Recommendations for mapping between a thesaurus and a classification scheme

The scope of a class should be determined by examining its caption(s), superordinate and subordinate classes, any descriptive notes and corresponding index entries.

The scope of each thesaurus concept should be established with similar care. See also Clause 14.

Mappings may be established from classes to thesaurus concepts, or from thesaurus concepts to classes, or both. It is important to take the direction of the mapping into account as work proceeds (see 13.2 for more discussion and examples).

For simple classes, it may be feasible to identify a single thesaurus concept that is exactly or inexactly equivalent (see 8.2 and Clause 11). Failing that, it may be possible to establish a hierarchical or associative mapping (see Clauses 9 and 10). But for classes involving the pre-coordination of simpler classes, typically the best solution available is a compound equivalence mapping (see 8.3).

For enumerative classification schemes, it may be sufficient to provide mappings for all the enumerated classes. For synthetic schemes (including faceted schemes), a comprehensive mapping capability should provide also for classes that are not enumerated but can be synthesized according to the rules of the scheme. The synthesized classes represent complex concepts. Typically they require compound equivalence mappings to the appropriate thesaurus concepts (see also 13.2.2).

While the hierarchical arrangement of the scheme may be found helpful in clarifying the scope of a given class, the inter-class relationships themselves need not be mapped (see definition of “mapping” at 3.45, which applies only to concepts and classes, not to relationships).

Mapping statements should be expressed using the notation to represent a class, rather than using its caption or corresponding index entries.

Further recommendations and examples, applicable to many taxonomies and to schemes used in records management as well as to classification schemes, are to be found in Clause 13.

18 Classification schemes used for records management

18.1 Key characteristics and background

18.1.1 General description

Many organizations use a custom-built classification scheme for the purpose of managing their official records. (It is sometimes known as a file plan, or as a business classification scheme.) While classification schemes used by libraries (see Clause 17) generally support access according to the *subject matter* of documents, the schemes used in records management systems are more concerned with *business context*. Consequently each record is typically associated with the business functions, activities and transactions it records, in the correct chronological series. A well designed scheme is usually based on analysis of the organization's activities, the regulatory environment and risk factors, and security and accountability requirements.

Figure 13 illustrates a typical functional classification scheme used for records management, showing only a few of the high-level class names and identifiers. All the classes shown are likely to be subdivided and subdivided again, to several more levels.

Identifier	Class name
HR000	Human resource management
HR100	Maintaining optimum levels of staffing
HR110	Staffing needs assessment
HR120	Recruitment
HR121	Advertising
HR122	Shortlisting
HR123	Interviewing
HR124	Informing applicants
HR200	Leave administration

Figure 13 — Extract from a functional classification scheme

Classification schemes for records management generally respond to legal requirements (which can vary from one jurisdiction to another) concerning the maintenance of complete and accurate records of the organization's activities. Retention and disposal requirements are a key design factor, because management is easier if classes are organized so that the entire contents of particular files and folders can be retained for their required period and disposed of according to a pre-determined schedule. Also the need for controlling access can influence the way files are grouped.

NOTE For further information, see ISO 15489-1 and ISO/TR 15489-2.

18.1.2 Scope and role in information retrieval

The classification scheme in a records management system typically supports browsing and has a limited role in searching.

As a browsing aid, the hierarchical structure of the scheme helps a user to find interconnected series of records at whatever level of aggregation is required. For example, the user might find a folder containing all the correspondence with American shareholders in a particular year and browse within it to find a subfolder for those in California only. Within that subfolder the user might pick out a file for one particular shareholder, and he might browse further within that file to identify a single record — perhaps an enquiry received from the shareholder.

When the names of files and folders in the scheme are known to the user of an electronic records management system, these can generally be used in search queries. But these names are typically inadequate when searching for records on a particular subject and so complementary search tools are needed, such as full text retrieval and/or indexing with a thesaurus.

18.1.3 Origin and development

In the days when most records were held on paper, the individual records were usually grouped in physical files, which were aggregated in folders and stored in filing cabinets. The files, folders and filing cabinets were labelled according to the classification scheme, which might or might not be listed in full in a separate set of schedules. With the advent of computers, such systems have been progressively automated, so that nowadays electronic systems are used for the vast majority of records. Increasingly these systems are being aligned with ISO 15489, and the practices of electronic records management systems are evolving rapidly. However, the analogy of files and folders, together with the tradition of hierarchical organization, still govern the classification schemes of most records management systems.

18.1.4 Vocabulary control

In records management systems, the names of classes in the classification scheme are often used in combination to compose the titles of individual records, as well as to label the files and folders. Different systems apply vocabulary control in different ways, but within any one system, naming consistency is vital.

When ISO 15489 is followed, the top level of the classification scheme generally reflects the principal business functions of the organization; the second level is based on the business activities constituting the function; the third and subsequent levels are further refinements of the activities or groups of transactions that take place within each activity.

To achieve unique identification of each class in the scheme, either its name is added to the names of its superordinate classes, or its code identifier or a combination of these can be used, depending on the rules of the system.

18.2 Semantic components and relationships compared with thesaurus components

18.2.1 Class names

Each class in the scheme usually has a name or label, comparable to the caption in a subject classification scheme (see 17.2.2). In different records management systems, the class names can be known as “headings” or “descriptors” or “keywords” or “titles” or “labels”. In any one array, each class name has to be unique, so that the user will not be confused. But in different arrays, it is common for the same name to recur. For example, “Bookings” might occur both under “Conferences” and under “Exhibitions”, as the name of two completely different classes. In the scheme shown in Figure 13, the class name “Advertising” might recur in the context of property sales.

18.2.2 Identifiers

Electronic records management systems generally allocate a unique identifier (such as a structured numeric or alphanumeric reference code) to each electronic class or file, right down to the level of individual records. The identifier is comparable with the notation of a subject classification scheme, except that it extends to the documents in the system, as well as the classes in the scheme.

A separate identifier is not always present in paper-based systems if the class names are judged sufficient for locating particular records. But where present, it often takes the form of a mnemonic code for each top-level class, followed by letters or numbers corresponding to each subdivision at successive levels. For example, the identifier “REG-PUB-COR” might represent the subfolder for “Correspondence” within the “Public consultation” folder of the top-level class “Regional planning”. An individual record within the subfolder might have identifier “REG-PUB-COR067”.

18.2.3 Other components of the scheme

Each class in the scheme typically has other data associated with it. The following elements are often found:

- description (comparable with the scope note of a thesaurus concept or class description of a subject classification scheme);
- index terms (including any mappings to or from thesaurus terms);
- dates (of opening, closing or deleting the file);
- access rights (specifying who is allowed access to the content);
- retention schedule (specifying how long the records in this file must be retained);
- history notes (comparable with the history notes of a thesaurus concept);
- cross-references to other classes in the scheme (comparable to “See also” references in a subject classification scheme, or the associative relationship in a thesaurus).

To encourage consistent application of the scheme, it is often accompanied by an index. Entries in the index can be derived from the name of the class and its superordinate classes, plus any useful terms in the description and their synonyms, plus any index terms assigned to the class.

Also accompanying the scheme there are often rules allowing administrators to extend the scheme, or to split or merge files, in the course of capturing new records and series of records.

18.2.4 Hierarchical structure

Like any classification scheme, that of a records management system involves pre-coordination of concepts (see 13.1) and it is usually monohierarchical. However, some systems allow a single record to be assigned to more than one class or file in the system.

18.3 Recommendations for interoperability with a thesaurus

The following potential use cases may be considered:

- a) A thesaurus may be used in the preparation of an index to the scheme. Although this indexing process could arguably be considered as a type of mapping, it will not be discussed further, since the resultant “mappings” are used internally within the scheme (usually to support the task of entering new records into the system) and not for external interoperability purposes.
- b) A thesaurus may also be used in the preparation of a subject index to the records stored in or managed by the records management system. This index is more extensive than that described in a), because individual records are indexed. A thesaurus used in this way complements the classification scheme, enabling searches on the basis of subject content rather than the business contexts addressed by the classification scheme.
- c) Another use case arises when documents from the records management system are selected for inclusion in another retrieval system, one which uses a different thesaurus. The cost of indexing with this thesaurus may be reduced by relying on mappings from the corresponding classes of the classification

scheme. The mappings allow appropriate thesaurus terms to be added to the metadata of the selected records [see also 12.1 a)].

- d) The use case described in c) could alternatively be addressed by mapping thesaurus terms to the classification scheme, and using the mappings to convert queries at the time of search [see also 12.1 b)].

Of these cases, only c) and d) involve mapping between the classification scheme and an external thesaurus. The recommendations and examples in Clause 13 provide general guidance, subject to the following additional recommendations.

- Before any mappings are prepared, the most appropriate way of representing each class in a mapping statement should be established, taking care that the chosen mode of identification is unique within the scheme.
- The direction of the mappings for use case c) is from the classification scheme to the thesaurus, with no need for reversibility. The mapping types described in Clauses 7 to 11 may be applied. However, cumulative compound equivalence should be applied with caution, as it could have a negative impact on the precision of searches.
- The direction of the mappings for use case d) is from the thesaurus to the classification scheme, with no need for reversibility. The mapping types described in Clauses 7 to 11 may be applied. However, intersecting compound equivalence should be applied only where the scheme is polyhierarchical and/or it is customary to assign more than one class to a single record.
- For both c) and d), it should be understood that both the recall and the precision that can be achieved by such mappings will be inferior to what can be done by indexing the individual records directly with the external thesaurus. This is because the classes of a records management system are rarely designed to reflect the subject content of records (see 18.1.1). For good retrieval results, a more effective approach is to index individual results as in b). If access by a second, different thesaurus is required, mappings between the two thesauri could provide an effective solution.

19 Taxonomies

19.1 Key characteristics and background

19.1.1 General description

The typical taxonomy is presented as a hierarchical vocabulary, used for classifying or categorizing, organizing, browsing, navigating, searching and/or filtering any type of content in networked environments. A common use case is to support navigation, especially by hierarchical organization and browsing through a broad set of electronic resources, e.g. websites, intranets, portals, wikis. Taxonomies are often used to provide website menus. To complement the navigational features with a search capability, taxonomies can include synonyms operating behind the scenes as entry terms, and “See also” references between related categories in the hierarchy.

Some taxonomies are customized to reflect the language, culture and goals of a particular organization, and used as the basis for knowledge sharing in the organization. They can be viewed as knowledge maps, or as communication and learning devices, integrating history, expertise and inside information in support of every business activity.

Some taxonomies include rules for automatic categorization and even personalization of incoming documents, for example news items. The rules may be used to sort incoming items into categories, and deliver them selectively to users expressing an interest in particular categories.

NOTE Automatic categorization and personalization techniques are outside the scope of this part of ISO 25964.

19.1.2 Scope and role in information retrieval

Taxonomies are commonly used for resource navigation in portals, intranets and websites, and they support retrieval primarily by enabling browsing.

To assist browsing, the categories in a taxonomy are generally arranged in hierarchies so that a user can navigate up or down to the appropriate level. Links between related categories in different hierarchies enable lateral browsing.

To assist searching, the categories have labels, often with multiple synonyms. The user can find what he wants by entering any of these directly.

To assist filtering, the categories where the search term appears can be presented, so that the user can choose to narrow the search to one or more categories.

An alternative approach to retrieval is to link a simple taxonomy with a more complex thesaurus. In this way the taxonomy can be presented very simply, as just a hierarchical set of categories and subcategories. Behind the scenes, however, can lie a much more elaborate thesaurus. If each category in the taxonomy is mapped to the corresponding thesaurus concept, all the preferred and non-preferred terms, as well as any scope note and relationships possessed by the concept, can be exploited to provide extra retrieval capabilities.

19.1.3 Origin and development

The term “taxonomy” originates from Greek, *taxis* (meaning “order” or “arrangement”) and *nomos* (meaning “law” or “science”). The original term is often used to refer to the classification and naming of organisms from an exclusive, monohierarchical perspective. In this part of ISO 25964, however, the definition of “taxonomy” is broadened to cover monohierarchical as well as polyhierarchical classification of any subject matter. In corporate settings, the taxonomy often includes entry points such as business functions, product types, business relationships, industry segments, business events, and document and record types.

19.1.4 Vocabulary control

Some taxonomies follow the model of classification schemes, providing a unique notation for each category. Much more commonly, however, the category label is not accompanied by notation. If the taxonomy follows the thesaurus model, each category label will be unique within the scheme and can be used alone to designate a particular category. Alternatively or additionally, a unique identifier can be provided, although generally it is hidden, intended mainly for computer use.

19.2 Types of taxonomies

19.2.1 General

The term “taxonomy” is used and misused so widely that many very different types of vocabulary can be given this name. In this clause, however, possible typologies will not be discussed because the focus is on taxonomies used for resource navigation. Even these include such a variety that in the context of interoperability it is useful to distinguish some key structural features that have an impact on mapping, as described in 19.2.2 and 19.2.3.

19.2.2 Monohierarchical versus polyhierarchical structures

For the purposes of interoperability, and especially the establishment of mappings, it is useful to distinguish taxonomies that follow the model of classification schemes from those designed more like thesauri. Figures 14 and 15 show contrasting examples.

Education	Health	Defence
Institutions	Institutions	Institutions
Funding	Funding	Funding
Research	Research	Research

**Figure 14 — Example of a simple monohierarchical taxonomy
(allowing repetition of category labels with differing scopes)**

Figure 14 is similar to the classification scheme illustrated in Figure 8 in Clause 13, except that there is no notation. Category labels are repeated in different hierarchies, but the categories are different. “Funding” in the education hierarchy refers only to the funding of educational activities; “Funding” in the health hierarchy to the funding of health activities, and so on. Just like the caption in a typical classification scheme, the category label is insufficient to identify the category uniquely, and should not be used alone in mapping statements.

Figure 15 shows a taxonomy in which each category label is unique, like the preferred terms of a thesaurus. The same category (e.g. “Military training”) can appear in more than one hierarchy; if so, the scope and content of the category is exactly the same, irrespective of its location. When unique category labels like these are available, they may reliably be used in mapping statements.

Despite some resemblances to thesauri, it should be noted that in Figure 15 the rules of hierarchy are not as strict as those of a true thesaurus (see ISO 25964-1:2011, 10.2). For example, a hierarchical relationship between “Education” and “Educational institutions” would not be recommended in a thesaurus.

Education	Health	Defence
Educational institutions	Health institutions	Defence institutions
Educational activities	Health-related activities	Military activities
Training	Medical training	Military operations
Medical training	Preventive medicine	Military training
Military training	Surgical operations	etc.
Teacher training	etc.	
etc.		

**Figure 15 — Example of a polyhierarchical taxonomy
(the same category can be included in more than one hierarchy)**

19.2.3 Non-symmetrical structures in multilingual and multicultural taxonomies

A great many taxonomies are developed for particular communities and for use by untrained laymen. This makes them subject to cultural and social influences that differ in parallel communities where a different language is spoken. For this reason non-symmetrical structures (as described in ISO 25964-1:2011, 10.2.6) are commonly found in multilingual taxonomies.

19.3 Semantic components and relationships

19.3.1 Categories

The basic units of a taxonomy, corresponding to the concepts of a thesaurus or the classes of a classification scheme, are usually known as “categories”. A category can encompass either a single concept, e.g. “Human rights” or a combination of concepts, e.g. “Government, citizens and rights”. A combination like this is often created as a presentational device to group several more specific categories. Sometimes the combination in a category is best conveyed by a phrase, e.g. “Coping with cancer” or “Search for clinical trials?”, corresponding to the anticipated needs of users. Sometimes the meaning of the category, single or combination of concepts, is expressed in imperative mood in order to communicate the sense persuasively (e.g. “Be safe in the sun” or

“Stay away from tobacco”). At lower levels of the taxonomy it is common to find a highly pre-coordinated expression such as “Car safety and child seats” or “Reporting crime and anti-social behaviour”.

In addition to the categories that apply to the subject or topic of content documents, taxonomies sometimes include non-subject categories such as:

- the intended audience (e.g. “Kids and Teens”);
- a time or place limitation (e.g. “Archive”);
- the form of documents (e.g. “Press releases”); and
- an associated task (e.g. “Claims submission”).

Any mix of subject categories with the above types can occur, at any level of the taxonomy.

Each category is given a category label, which in some taxonomies is unique but in others is not disambiguated (compare Figures 14 and 15). Where disambiguation is required, the label is sometimes given a qualifier (see ISO 25964-1:2011, 6.2.2). The category labels correspond to preferred terms in a thesaurus. In a multilingual taxonomy, each category usually has a different label for each language.

As well as the label for human use, each category in some taxonomies also has a unique, non-semantic identifier for machine use. This is especially useful when category labels are not disambiguated. Provision of a user-readable notation is unusual.

19.3.2 Synonym variations

Category labels often have synonyms in the same way as the preferred terms in a thesaurus have non-preferred terms. Like the terms in a synonym ring, these can include near-synonyms, abbreviations, acronyms and lexical variants (see 24.2). Qualifiers are sometimes added for disambiguation. The equivalence relationship between a category label and its synonyms can be displayed explicitly, but more often it is hidden. An alternative way of invoking synonyms is to provide mappings from categories to the corresponding entries in a thesaurus or set of synonym rings; although the synonyms are not present in the taxonomy itself, they can be used by invoking the mappings at the time of search.

19.3.3 Hierarchical relationships

In some taxonomies the rules governing hierarchical relationships in thesauri are followed (see ISO 25964-1:2011, 10.2), including the use of polyhierarchical structures (see 19.2.2). More commonly, however, taxonomy hierarchies are like those of a classification scheme (see 17.2.3) and for these the structure is usually monohierarchical (see 19.2.2), with implicit pre-coordination of concepts. See Clause 13 for more discussion of how to handle pre-coordination.

19.3.4 Associative relationships

Some taxonomies incorporate associative relationships between related categories, usually from different hierarchies. Comparable to the associative relationship in a thesaurus (see ISO 25964-1:2011, 10.3), they can appear as “See also” references, in the style of many classification schemes, and/or be implemented as hyperlinks between the related categories.

19.3.5 Definitions and scope notes

A definition and/or scope note may be attached to a category where necessary to clarify its scope.

19.3.6 Rules for automatic categorization

Optionally, rules for automatic categorization, automatic querying and personalization can be attached to the categories.

19.4 Mapping between a thesaurus and a taxonomy

19.4.1 General

In common usage the term "taxonomy" is applied widely, with a great variety of different meanings. Therefore, before mappings are developed to or from a vocabulary described as a "taxonomy", the vocabulary should be studied to identify its type and principal characteristics. Characteristics which can influence the content and style of mappings include:

- presence/absence of pre-coordination (see Clause 13, especially Figure 8);
- presence/absence of polyhierarchy;
- presence/absence of non-symmetrical structures (in a multilingual taxonomy);
- availability of notation;
- availability of a unique category label functioning as preferred term; and
- inclusion of proper names.

Where pre-coordination in the style of Figure 8 is found, the taxonomy can be regarded as a type of classification scheme and the recommendations in Clause 13 and 17.3 should be followed. If, however, each category label is unique in the taxonomy, and especially where polyhierarchical structures are permitted, the taxonomy can be treated more like a thesaurus and pre-coordination between subordinate and superordinate categories is less likely to cause difficulties.

Some use cases require mappings from a thesaurus to a taxonomy, while others need mappings in the opposite direction. Either option is in general feasible and the examples in Clauses 12 and 13 may help guide the choice.

All of the mapping types described in Clauses 7 to 11 may be considered.

A compound equivalence mapping is usually needed when mapping from a grouping category such as "Government, citizens and rights", or "Accidents and accident prevention".

In all mapping statements, the category should be designated unambiguously, using the category label if it is unique, or a non-semantic unique identifier. The former designation is generally more appropriate for human readers; the latter for manipulation by computers.

In the case of a non-symmetrical multilingual taxonomy, a choice should be made concerning which language version or versions to use in the mappings. In the symmetrical parts of the taxonomy, a mapping to or from any category should be equally valid in all its languages. In the non-symmetrical parts, to achieve comprehensive mapping each category in each language version should be mapped separately. In some contexts, it may be sufficient to prepare mappings for only one of the language versions. But if this is done, the mappings should not be applied to contexts using the other languages.

In the case of non-subject categories, it is difficult to develop effective mappings, because thesauri designed for subject indexing do not usually contain the terms or concepts needed for indexing other metadata elements such as document type or audience. If the thesaurus is not used as the source of values for these elements, mappings between taxonomy and thesaurus should not be made. For a particular application, it is sometimes possible to develop mappings from the non-subject categories to a combination of values drawn partly from the thesaurus and partly from the authority list used for the metadata element in question (see Example 7 in 19.4.2 below).

19.4.2 Practical examples

In all the mapping statements for the following examples, the identifier of the thesaurus is "THES" while that of the taxonomy is "TAX". For purposes of these examples it is assumed that the category label is unique; if this is not the case some other, unambiguous, identifier should replace it in the mapping statements.

EXAMPLE 1

Mapping case	Taxonomy has a category “rights”, while the nearest concept in the thesaurus is “civil rights”.
Mapping statement	rights THES NM civil rights
Discussion	In this case, mapping from taxonomy to thesaurus is much the same as between two thesauri. See Clauses 7 to 10.
Mapping in the reverse direction	civil rights TAX BM rights

EXAMPLE 2

Mapping case	Taxonomy has a category “human rights”, while the nearest concept in the thesaurus is “civil rights”.
Mapping statement	human rights THES ~EQ civil rights or human rights THES NM civil rights
Discussion	Which of these mappings is more accurate depends on the context in which the mappings will be used; for example, the most applicable legal jurisdiction under which the rights are defined.
Mapping in the reverse direction	civil rights TAX ~EQ human rights or civil rights TAX BM human rights

EXAMPLE 3

Mapping case	Taxonomy has a category “jams, jellies and preserves”, while the nearest concept in the thesaurus is “fruit preserves”.
Mapping statement	jams, jellies and preserves THES EQ fruit preserves
Discussion	Possibly, the statement could be modified to mark it either exact or inexact. To determine which of these is appropriate, it is necessary to study any scope notes associated with the concepts, the hierarchical environment of the category, and the way resources have been assigned to the category/concept.
Mapping in the reverse direction	fruit preserves TAX EQ jams, jellies and preserves

EXAMPLE 4

Mapping case	Taxonomy has a category “children and obesity”, while the thesaurus has separate concepts “children” and “obesity”.
Mapping statement	children and obesity THES EQ children + obesity
Discussion	Although the “and” makes this look similar to Example 4, it is in fact an example of intersecting compound equivalence and should be handled differently. The category is not intended to group everything about children and everything about obesity, but only for items dealing with the combination of concepts.
Mapping in the reverse direction	The following separate statements should be considered: children TAX NM children and obesity obesity TAX NM children and obesity Before either of these is accepted, the taxonomy should be checked, to see if there are other categories that match the thesaurus more closely, or perhaps equally closely. Multiple mappings from the same concepts could well be appropriate.

EXAMPLE 5

Mapping case	Taxonomy has a category “government, citizens and rights”, and there is no single concept in the thesaurus that brings all this subject matter together. However, the thesaurus does have separate concepts “government”, “general public” and “civil rights” as well as several other rights-related concepts.
Mapping statement candidates	a) government, citizens and rights THES EQ government general public civil rights b) government, citizens and rights THES EQ (government civil rights) + general public c) government, citizens and rights THES EQ government civil rights d) government, citizens and rights THES EQ government civil rights consumer protection e) [no mapping]
Discussion	At first sight this looks like an example of cumulative compound equivalence, as shown in option a). However, option a) is misleading because this category does not really include information <i>about</i> citizens or the public. The mention of citizens is intended to suggest that the information is <i>for</i> citizens (rather than <i>for</i> businesses or <i>for</i> the state). Option b) looks closer, because it takes in information about government and/or civil rights, which also mentions the general public. The weakness of option b), if applied to a search expression, is in requiring the presence of “general public” among the metadata of resources being searched. Even when a document is written expressly for the guidance of the citizen and/or the general public, this aspect might commonly not be captured in its metadata. To avoid excluding relevant material, option c) performs better (though such an open-ended search statement might not yield good precision). Option d) could result from examining the taxonomy more closely and discovering “consumer rights” among the subcategories of “government, citizens and rights”. Consumer rights are not usually considered among civil rights, and so more concepts need to be added to the mapping statement, making it more open-ended than before. Plainly all the proposed mappings are inexact, but a marker is not needed because all compound equivalence mappings are inexact. In this case, the mappings seem so very inexact that one should question whether any mapping is really needed. Option e) seems appropriate if the mapping has no important use. Plainly the category “government, citizens and rights” is useful in the taxonomy as a presentation device to assist users in navigation down to subcategories such as “citizenship”, “identity cards”, “consumer rights”, etc. but it is much less useful in search statements or in document metadata. A mapping of a grouping category like this can often be disregarded if a thorough job is made of mapping all its subcategories.
Mapping in the reverse direction	A single mapping is not available. But the following separate statements should be considered: a) government TAX BM government, citizens and rights b) general public TAX BM government, citizens and rights c) civil rights TAX BM government, citizens and rights d) consumer protection TAX BM government, citizens and rights Option g) is inappropriate, because the category is not <i>about</i> citizens or the public. Options f), h) and j) could be valid, but before any of these is accepted, the narrower categories of “government, citizens and rights” should be checked. Among them the taxonomy may well have categories (such as “consumer rights”) that correspond much more closely to the thesaurus concepts than does the broader category. Mappings at the same level of specificity are usually preferable to hierarchical mappings and so, for example, “consumer protection TAX ~EQ consumer rights” would be preferred to “consumer protection TAX BM government, citizens and rights”.

EXAMPLE 6

Mapping case	Taxonomy has a category "Coping with cancer", and there is no single concept in the thesaurus that combines "coping" and "cancer". The thesaurus does have the concept of "cancer". But the nearest single term/concept to "coping" is "management" or "disease control" or "medical treatment" or "health protection", none of which seem to capture what is intended in the taxonomy. The category has much appeal for a community of cancer sufferers, carers and specialists, but it is made up of very many aspects, such as therapy, palliative care, counselling, childcare, medication, lifestyle, etc.
Mapping statement candidates	a) Coping with cancer THES EQ cancer + (support services medical treatment complementary medicine palliative care counselling childcare medication lifestyle respite) b) Coping with cancer THES RM cancer
Discussion	Option a) seeks to capture the whole of what is implicit in "coping", but the outcome is far from perfect. A statement like this could not be used for conversion of index terms (except in a prompting system for human indexers). If used for conversion of search statements, high recall and precision cannot be expected. Option b) acknowledges the difficulty of capturing the whole essence of "coping". Again it could be useful as a prompt for a human searcher or indexer, but is not suitable for automatic conversion of search statements or index terms.
Mapping in the reverse direction	The following statement should be considered: cancer TAX RM coping with cancer Before the mapping is accepted, it should be checked whether the taxonomy has a narrower category corresponding more closely to "cancer". If it is available, an equivalence mapping is generally preferable to hierarchical or associative mappings. Statements such as the following can be considered too: support services TAX RM coping with cancer medical treatment TAX RM coping with cancer etc. In each case, the taxonomy should be inspected for the presence of other, more closely matching categories, and the closest should be preferred.

EXAMPLE 7

Mapping case	Taxonomy for a website dealing with health care has a top-level category "Kids and Teens", while the thesaurus has concepts "children" and "teenagers", as well as "health care". But in systems and collections using this thesaurus, the source of terms used for the metadata element "Audience" is a small authority list called "Targets" (abbreviated TARG), which contains the values "Children" and "Youths".
Mapping statement candidates	a) Kids and Teens THES BM health care b) Kids and Teens TARG BM Children Youths c) Kids and Teens EQ TARG (Children Youths) + THES health care
Discussion	The taxonomy category labelled "Kids and Teens" is narrower than the label implies, being in fact "Health care for a target audience of children and teenagers". The label is misleading when taken out of context. A mapping from this category to thesaurus concepts "children" and "teenagers" is inappropriate, because it would have the effect of searching for documents <i>about</i> children, instead of documents written <i>for</i> children. It would be similarly misleading if applied to the conversion of index terms. Rather than provide no mapping at all, option a) is a possibility, although it will not discriminate between audiences. The tag BM (rather than EQ) should be used, because "health care" is the broader concept applicable to the entire content of the website. Option b) is available only if the application makes provision for directing the search (or converted index terms) to the Audience metadata field rather than the Subject field. It too uses tag BM rather than EQ, because the search would retrieve <i>all</i> material intended for a young audience, not just items relevant to health care. Option c) is potentially more powerful, combining the desired subject area with the Audience restriction. But the syntax it uses is not standard across external search systems. The results are likely to be effective only in a purpose-built application within the control of the systems engineers.
Mapping in the reverse direction	There are risks associated with applying any of these options in the opposite direction. At most, they could be used in a "suggester" interface that offers candidate terms to the indexer or searcher.

20 Subject heading schemes

20.1 Key characteristics and background

20.1.1 General description

A subject heading scheme is a type of controlled vocabulary that is used to represent in synthesized form the topics discussed in documents of any type. Subject heading schemes share some characteristics with thesauri and with classification schemes. Like a thesaurus, they represent concepts in the form of terms or phrases, and like a synthetic classification scheme, they provide syntactic rules for combining terms into pre-coordinated strings that represent more complex concepts and topics. Well-known examples of subject heading schemes are the *Library of Congress Subject Headings* (LCSH)^[9] and *Medical Subject Headings* (MeSH)^[24] schemes.

20.1.2 Scope and role in information retrieval

A primary function of subject heading schemes is to bring together related topics so that collections of information resources can be organized systematically on the basis of their subject content, and to facilitate browsing and navigation around a subject domain.

At the time of indexing, the indexer is responsible for putting together all the components that are necessary to represent all facets of a topic, thus creating a more or less complex subject heading. At the time of searching, where the software offers these functionalities, the end-user can retrieve information resources by using any single word appearing in the subject heading, any combination of words appearing in the subject heading, or the subject heading as a whole. The use of single words and of any combination of words will favour recall, while precision will be increased when the subject heading as a whole is used as search query.

20.1.3 Origin and development

Subject heading schemes were created at the end of the 19th century as a tool used to systematize subject access to information resources, when cataloguers began to create lists of subject terms to ensure indexing consistency within their own institution's catalogue. Since then, various lists of subject headings have been established and used. The *Library of Congress Subject Headings* (LCSH) remains the most widely used subject heading scheme, not only in the English-speaking world, but also, through translations and adaptations, in a large number of institutions whose working language is not English.

20.1.4 Vocabulary control

Just as a thesaurus guides indexers and searchers towards consistent use of the same term(s) to represent a given concept, similarly a subject heading scheme encourages consistent use of the same heading(s) to represent the same concept. Thus it differs from a classification scheme, which uses a notation for this purpose.

20.1.5 Types of subject heading schemes

Encyclopaedic or general subject heading schemes cover all disciplines and are most commonly used to index and retrieve information resources in general collections. LCSH^[23], *Répertoire des vedettes-matière de l'Université Laval* (RVM)^[28], *RAMEAU*^[26], *Nuovo Soggetario*^[10] and SWD (Schlagwortnormdatei), which now forms part of the GND (Gemeinsame Normdatei)^[29], are examples of general subject heading schemes. Disciplinary or specialized schemes cover a single topic or discipline (for example, *MeSH*^[24] and *Subject Headings for Music* (SHM)^[12]) or a specific category of information resources (for example, the *LC's Thesaurus for Graphic Materials* (TGM)^[22]). Some schemes are appropriate for a particular category of end-users (for example *LC's Subject Headings for Children*)^[34].

20.2 Semantic components and relationships

20.2.1 Overview

The main components of a subject heading scheme are headings, subdivisions and relationships between them. The scheme may also include or be accompanied by rules for when and how to combine these components.

20.2.2 Headings

Headings can consist of a word representing a single concept (e.g. "Arithmetic"), of several words representing one concept (e.g. "Arithmetical algebraic geometry") or of several words representing a combination of distinct concepts (e.g. "Mines and mineral resources"). The form and meaning of each heading are controlled, albeit less strictly than is the case in a thesaurus. For example, the plural form of count nouns is normally preferred in English versions of subject heading schemes (e.g. "Trees"; "Forests"; "Arid regions"). In multiword headings, words can be presented in direct, natural order (e.g. "Dacians in literature"; "River sardine fisheries") or in inverted order (e.g. "Civilization, Dacian"; "Series, Arithmetic"). Homographs are disambiguated by a qualifier, e.g. "Mercury (Planet)" and "Mercury (Roman deity)".

Subject headings can be usable for indexing (accepted headings) or not usable for indexing (rejected headings). These are comparable to preferred and non-preferred terms, respectively, in a thesaurus.

A subject heading can be simple, consisting of just one concept, or it can be compound, if one or more subdivisions are added to the initial heading to form a pre-coordinated string.

20.2.3 Subdivisions

Compound headings are pre-coordinated strings created by adding to the initial heading one or more subdivisions to represent a topic completely and accurately. A double dash generally connects the subdivision to its initial heading, as shown in Figure 16. Subdivisions can consist of a single word, e.g. "Marketing" or of several words e.g. "Meiji Period, 1868-1912". The function of subdivisions is to represent the perspective, point of view, form, etc. under which the topic expressed by the heading is presented in an information resource.

Common types of subdivisions include:

- a) topical subdivisions, e.g. "Marketing", "History", "Alternative treatments", "Colonies", "Genetic aspects";
- b) geographical (place) subdivisions, e.g. "Italy", "Russia", "Québec (Province)";
- c) chronological (time) subdivisions, e.g. "500-1400", "20th century", "Meiji period, 1868-1912"; and
- d) form subdivisions, e.g. "Dictionaries", "Directories", "Statistics", "Amateurs' manuals".

Many subdivisions are to be used only with specific headings or with certain categories of headings e.g. "Storage — Diseases and injuries: Use as topical subdivision under individual plants and groups of plants". A few subdivisions, known as free-floating subdivisions, can be added to any or most headings, e.g. "Pictorial works: Use as a form subdivision under names of countries, cities, etc., individual persons, families, and corporate bodies and other named entities, such as individual parks, structures, etc., and under classes of persons, ethnic groups, individual wars, and topical headings".

When more than one subdivision is appended to a heading, they normally appear in a consistent order, such as: — Topical — Geographical — Chronological — Form.

20.2.4 Relationships

The three main types of relationships found in thesauri are also used to structure most subject heading schemes. They are:

- a) the equivalence relationship (indicated by cross-references such as See, See from, or by the tags USE and UF);
- b) the hierarchical relationship (indicated by level of indentation, by typographic devices, or by explicit tags such as BT and NT); and
- c) the associative relationship (indicated by cross-references such as See also or by the tag RT).

The nature and functions of these relationships in subject heading schemes are as described in ISO 25964-1:2011, Clauses 9 to 11.

20.3 Mapping subject headings to and from thesaurus concepts

If a subject heading scheme lists only simple headings, without enumerating any pre-coordinated strings, mapping between this scheme and a thesaurus should follow the same guidelines as for mapping between two thesauri (see Clauses 7 to 11). Extra guidance is needed, however, when compound headings are handled.

To illustrate this section, we shall use examples drawn from a subject heading scheme enumerating the headings and strings shown in Figure 16.

automobiles
automobiles — brakes
automobiles — clutches
automobiles — clutches — handbooks
automobiles — clutches — maintenance — handbooks
automobiles — handbooks
automobiles — maintenance
automobiles — maintenance — handbooks
brakes
clutches — maintenance
maintenance
washing machines — maintenance — handbooks

Figure 16 — Some enumerated headings in a subject heading scheme

20.3.1 Mapping from a subject heading scheme to a thesaurus

Three approaches should be considered when mapping from a subject heading scheme to a thesaurus. An approach should be chosen in the light of the context of the application, taking into account the arguments that follow.

- a) Map separately each heading and subdivision of the simple and compound headings enumerated in the source vocabulary, to the target thesaurus.

EXAMPLE 1

The subject heading “automobiles — maintenance” is not mapped as a whole, but “automobiles” and “maintenance” are each mapped independently to the target thesaurus.

This approach achieves a consistent and comprehensive mapping of the constituent elementary concepts, without requiring the syntactic rules of the subject heading scheme to be taken into account. As these rules can be complex, they might not be easy to apply in an automated system. In this approach, however, complex concepts expressed as pre-coordinated strings in the source subject heading scheme are not mapped, even when there is a corresponding complex concept in the target thesaurus. This can result in a loss of precision in retrieval, if the mappings are used to convert index terms and strings.

EXAMPLE 2

Even though the compound heading “automobiles — maintenance” has no direct mapping, when found as an indexing string it can be converted to a combination of “automobiles” and “maintenance”. This is unsatisfactory, however, if the target thesaurus incorporates the complex concept “automobile maintenance”. A user searching with this thesaurus term will not retrieve relevant items that were indexed with the compound heading.

- b) In addition to a), also map all the compound headings that are enumerated in the source subject heading scheme. If possible, map each one to a pre-coordinated concept in the target thesaurus. If a pre-coordinated concept is not available in the target thesaurus, map the source string to a combination of target concepts. See further guidance on compound equivalence in 8.3.

EXAMPLE 3

The subject heading “automobiles — maintenance” is mapped to the target concept “automobile maintenance” if it exists in the thesaurus. If this does not exist in the thesaurus, the pre-coordinated string can be mapped to the combination “automobiles + maintenance”, as described in 8.3.2.

This extends the number of complex concepts that are mapped, but it might lead to inconsistencies and uncertainty. The subject heading scheme might enumerate the string “automobiles — maintenance”, with a rule saying that the topical subdivision “maintenance” can be used with any heading representing a type of land vehicle. Therefore, the string “bicycles — maintenance”, although valid, would not be mapped if only enumerated strings were mapped to the target thesaurus.

The limitations of both a) and b) are difficult to overcome without some human intervention and interpretation at the time of using the mappings, either for conversion of metadata or for conversion of search queries.

- c) Map only the subject headings that have actually been used to index a collection of information resources, rather than the whole subject heading scheme. This will have the effect of mapping pre-coordinated strings not enumerated in the source vocabulary, while omitting enumerated strings that have not yet been used as indexing terms in a local collection of resources.

This approach offers the following advantages:

- human and financial resources do not need to be assigned to the mapping of subject headings and subdivisions that are unlikely ever to be used for indexing;
- the fact that all and only those subject headings that have actually been used for indexing are mapped to the target thesaurus improves both precision and recall at the time of retrieval.

Its disadvantages are:

- since compound subject headings constructed by indexers will include many of the same constituent elements in different combinations, the workload increases and there is a danger of providing conflicting mappings for the same terms (repeated mapping of the same constituent should be avoided);

- the result of the mapping process is valid only in the context of a specific catalogue, index or database, and may not be valid for other catalogues, indexes or databases even if they use the same source subject heading scheme as indexing vocabulary.

Once the complete set of mappings has been prepared by any of the above approaches, it can be used either to convert indexing terms or to convert search terms (see more discussion in 12.1). If indexing terms are converted, a thesaurus user will be enabled to access a collection indexed with the subject heading scheme. If search terms are converted, a user of the subject heading scheme will be enabled to access a collection indexed with the thesaurus.

20.3.2 Mapping from a thesaurus to a subject heading scheme

When mapping from a thesaurus to a subject heading scheme, the recommended approach is to map all the concepts in the thesaurus, following the guidelines in Clauses 7 to 12. If resources are scarce, economies may be achieved by mapping only those concepts that have actually been used as index terms in a particular catalogue, index or database. This limitation, however, will cause problems when documents covering new subjects are added to the collection and indexed with terms not previously used.

Any complex concepts in the source thesaurus should be mapped to enumerated headings or strings if available. If a suitable pre-coordinated string is not enumerated in the target subject heading scheme, it should be created by combining an appropriate heading with one or more subdivisions according to the syntactic rules provided in the scheme or in accompanying documentation.

EXAMPLE 1

The source concept “automotive engineering” can be mapped to the subject heading “automobiles — design and construction”, if the rules of the target subject heading scheme permit the creation of this pre-coordinated string.

Once the complete set of mappings has been prepared, it can be used either to convert indexing terms or to convert search terms (see more discussion in 12.1). If indexing terms are converted, a user of the subject heading scheme will be enabled to access a collection indexed with the thesaurus. If search terms are converted, a thesaurus user will be enabled to access a collection indexed with the subject heading scheme.

Automated synthesis of subject headings on the basis of the co-occurrence of terms in the metadata of a particular document is not recommended, without human validation or a purpose-built algorithm. For example, if the index terms “handbooks”, “training”, “automobiles”, “clutches”, “gears”, “cleaning” and “costs” have all been assigned to a document, at least seven of the subject headings listed in Figure 16 are potentially applicable, as well as many others. Without inspecting the document, it is hard to know which of these combinations are appropriate. In this case, mapping should be based on the concepts actually found in the thesaurus, not on combinations of them.

EXAMPLE 2

If the thesaurus concepts “automobile maintenance” and “costs” have been mapped to “automobiles — maintenance” and “costs” respectively, the search expression “automobile maintenance AND costs” can be converted to “(automobiles — maintenance) AND costs”.

Alternatively, if the index terms “automobile maintenance” and “costs” occur in the metadata of a document, they can be converted to the subject headings “automobiles — maintenance” and “costs”, but without inspection of the document an indexing string such as “automobiles — maintenance — costs” should not be concocted.

21 Ontologies

21.1 Key characteristics and background

21.1.1 General description

The term “ontology” has many different meanings, ranging from semantic data models to the more philosophically influenced application of categorial and metaphysical analysis of the concepts in a domain. In computer science, more specifically in the field of knowledge engineering and artificial intelligence, an ontology is described as “a formal, explicit specification of a shared conceptualization”. This extension by Studer et al.^[31] of Gruber’s original formulation in Reference [18] is the definition used in this part of ISO 25964. In the context of data modelling, “ontology” is often interpreted as the use of a formal language to set out a formalized representation of a domain of knowledge. Among other tasks, this enables the consistency of knowledge assertions (facts) to be checked against the ontology, and possibly new ones to be inferred. An ontology and a set of facts (assertions about individuals) together form a knowledge base.

One of the fundamental purposes of an ontology is reasoning, including generic tasks such as:

- inferring class membership for individuals;
- inferring relationships between classes and properties; and
- checking the consistency of a knowledge base.

EXAMPLE

In the medical field, for example, an ontology of diseases might:

- use measurements of the patient’s symptoms to infer which disease is involved; or
- discover that the different symptoms presented by two patients were both caused by the same virus.

21.1.2 Scope and role in information retrieval

Whereas the role of most of the vocabularies described in this part of ISO 25964 is to guide the selection of search/indexing terms, or the browsing of organized document collections, the purpose of ontologies in the context of retrieval is different. Ontologies are not designed for information retrieval by index terms or class notation, but for making assertions about individuals, e.g. about real persons or abstract things such as a process. While internal navigability is not usually of prime concern, ontologies can have a role in some of the retrieval use cases described in 21.4.

21.1.3 Origin and development

The term “ontology” has become popular through its application in the field of computer science (particularly artificial intelligence and knowledge engineering) from the early 1990s to mean the formal specification of a body of knowledge that can be used for reasoning. Since then, much research effort has been spent on the development of formal languages for use in describing ontologies and in enabling automated reasoning (see also 21.1.6).

In the biomedical domain, some complex ontologies are being integrated with other ontologies in the same or neighbouring domains and still provide results in logic-based reasoning that have correct and useful correspondence to real life^[29]. In this domain an ontology, more than any other type of vocabulary, requires careful analysis of the nature of the real-life entities represented in it. Thus ontologies as computer science artefacts regain some of the original context of the term “ontology”, a branch of philosophy also known as general metaphysics, or the science of being.

More recently, the term “lightweight ontology” has been employed in some Semantic Web literature to cover all sorts of structured vocabularies and knowledge organization systems, including thesauri, classification

schemes, etc. This terminology is not employed in this part of ISO 25964, since the blurring of distinctions entailed in the loose use of the term is considered unhelpful. For guidance on the use of SKOS (Simple Knowledge Organization Systems, see Reference [37]) to publish and link thesauri (and other knowledge organization systems) on the Semantic Web, see ISO 25964-1:2011, Clause 17.

21.1.4 Vocabulary control

Vocabulary control (as defined in 3.98) is not the specific aim of ontologies. For example, they do not always take care to disambiguate homographs, as would be expected in thesauri. In ontologies the disambiguation of terms is less important, since both classes and individuals are uniquely identified by means other than natural language. However, the usefulness of unambiguous labelling is often acknowledged in the ontology research literature.

21.1.5 Types of ontologies

Ontologies can be distinguished by (for example) the degree of specificity, the domain covered or the application purpose^[19]. Thus top-level ontologies distinguish the most general, domain-independent categories of existence, while domain ontologies and task ontologies describe, respectively, the classes in a specific domain and those for a specific task, sometimes based on top-level ontologies.

While these distinctions are not clear cut, the ontologies covered in this part of ISO 25964 are limited to formal domain ontologies having complexity and coverage comparable to those of the structured vocabularies typically used for information retrieval.

21.1.6 Logics and languages for expressing ontologies

Generally, first-order logic (or a subset) is used for expressing ontologies. A variety of representation languages are available for the description of ontologies. Examples are the Resource Description Framework Schema (RDFS)^[36], or the Web Ontology Language (OWL)^[35] which are recommendations of the World Wide Web Consortium (W3C). OWL provides a set of axioms, which is deliberately limited so that tractable algorithms for reasoning can be provided.

21.2 Semantic components and relationships

21.2.1 Overview

This discussion of semantic components adopts OWL terminology. The main components to consider are classes, properties, axioms and individuals (instances). Some other features of ontologies are also described briefly in this clause.

NOTE The illustrative example in 21.2.9 uses the CamelCase naming convention. UpperCamelCase notation is used to denote identifiers of classes, while lowerCamelCase is used for relation identifiers. For human readability, natural language labels could be additionally provided according to a specific labelling convention.

21.2.2 Classes

An ontology class is a construct with a set of property constraints that establish the criteria for membership. In contrast to classification schemes, ontologies use a formal language to express the properties that serve to define classes explicitly. Thus an ontology can define classes in relation to other classes using logical connectors or other restrictions. For example a class can be defined as the intersection between two classes (denoting all individuals that belong to both classes), or as the complement of another class (denoting the individuals not belonging to another class).

21.2.3 Properties

Each class can be described by properties of its members (individuals). Properties can be either:

- attributes (e.g. `hasTitle`, `hasMass`) to which particular values can be assigned (e.g. 50 grams); or
- relationships between members of one class and those of other classes (e.g. the relation “`planetOrbits`” in “Earth `planetOrbits` Sun”; see the example in 21.2.9).

The way in which properties are used can be subject to restrictions, e.g. the class “Planet” in the example in 21.2.9 might be specified to have only a single value as its “`hasMass`” attribute.

21.2.4 Axioms

Axioms are statements specifying the basis of the classes, properties and other entities in an ontology. An axiom can be as simple as a statement that a specific class or property exists. Other examples of axiomatic statements describe the properties of a class (e.g. the class “Planet” is required to have an attribute “`hasMass`”).

21.2.5 Class hierarchies

Hierarchical structure in ontologies is created through sub-class axioms between classes. The sub-class relation implies that all axiomatic statements of a parent class also apply to its child classes, such that the property constraints are inherited by all of the children. The sub-class relation is transitive, i.e. it propagates downwards to all sub-classes.

21.2.6 Individuals

Individuals are the objects of discourse in a particular domain and are sometimes referred to as instances. They are the fundamental things that an ontology makes statements about. Examples of individuals are a specific person, a particular printed book, but also abstract things, such as a particular symptom of a clinical patient, a particular process or event, or a particular occurrence of a feeling such as love.

21.2.7 Assertions

Assertions, a particular group of axioms, are statements about individuals in a domain. The “class assertion”, in particular, states that an individual is a member of a class. Other assertions could state that two individuals are the same or that they have certain property constraints. Assertions and individuals are generally brought in as part of the ontology application rather than including them in the ontology itself. Nonetheless both can be described by the ontology language OWL, and the assertions need to comply with the other axioms in the ontology.

21.2.8 Labels and identifiers

Classes and properties, as well as individuals, all have identifiers. Although natural language labels are not strictly necessary in ontologies, they are often provided to enhance readability for humans.

21.2.9 Illustration of a simple ontology

Tables 13 to 15 show some key features of ontologies. A graphical illustration appears in Figure 17.

NOTE This example is highly simplified in order to show some key features of ontologies in comparison to thesauri.

Table 13 — Classes of an ontology snippet in the domain of astronomy

Class name	Properties/Axioms
AstronomicalObject	hasMass: datatype
Planet	subClassOf: AstronomicalObject planetOrbits: <i>some</i> Star
Star	subClassOf: AstronomicalObject

Table 14 — Some properties of an ontology snippet in the domain of astronomy

Property name	Axioms
<i>hasMass</i>	Domain: AstronomicalObject Range: Mass datatype (in kg)
<i>orbits</i>	Domain: AstronomicalObject Range: AstronomicalObject InverseOf: <i>isOrbitedBy</i>
<i>planetOrbits</i>	Domain: Planet Range: Star InverseOf: <i>isOrbitedByPlanet</i>
<i>isOrbitedBy</i>	Domain: AstronomicalObject Range: AstronomicalObject inverseOf: <i>orbits</i>
<i>isOrbitedByPlanet</i>	Domain: Star Range: Planet inverseOf: <i>planetOrbits</i>

Table 15 — Examples of individuals (instances) in the domain of astronomy

Instance name	Assertions
Earth	instanceOf: Planet planetOrbits: Sun hasMass: $5,974 \times 10^{24}$ kg
Sun	instanceOf: Star isOrbitedByPlanet: Earth hasMass: $1,989 \times 10^{30}$ kg

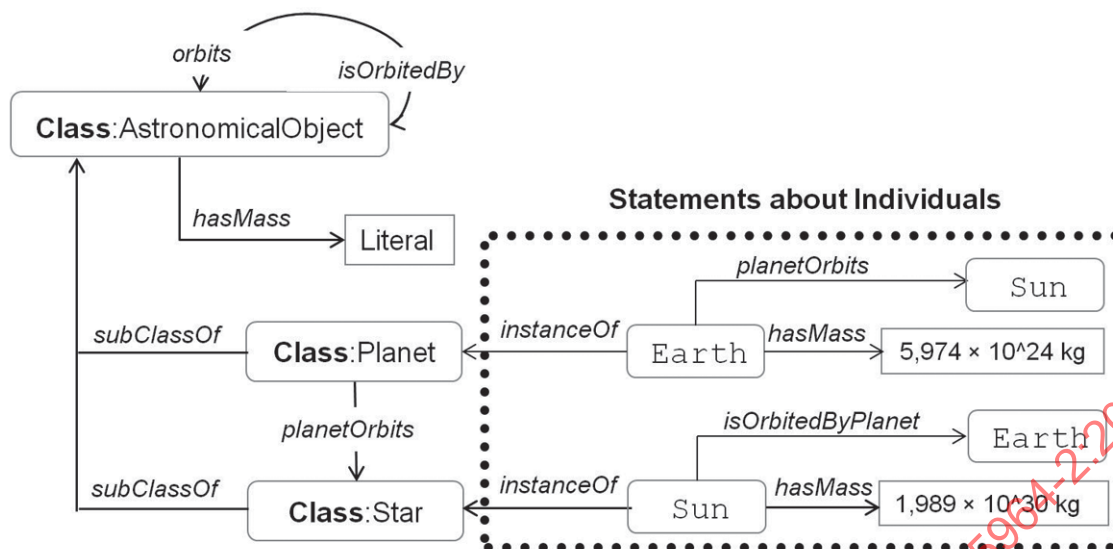


Figure 17 — Illustration of an ontology snippet and assertions

In this example, all sub-classes of “AstronomicalObject”, such as “Planet” or “Star”, inherit all properties of their super-class. This ontology states that every astronomical object is required to have a mass property and some astronomical objects can orbit another astronomical object. Since “Planet” is a sub-class of “AstronomicalObject”, it can be inferred that a planet has a mass and can orbit another astronomical object. An ontology further allows constraints on the use of properties. Here, by introducing a sub-property, “planetOrbits”, the ontology requires all planets to orbit a star.

21.3 Structural comparison between thesauri and ontologies

One key difference is that, unlike thesauri, ontologies necessarily distinguish between classes and individuals, in order to enable reasoning and inferencing.

EXAMPLE

“Earth” and “planets” might be two concepts in a thesaurus, linked by a hierarchical BT/NT relationship. Optionally, the instantial nature of the relationship could be made explicit by using the tags BTI and NTI.

In an ontology concerned with celestial mechanics, “Planet” could be established as a class, and “Earth” as an individual (since the latter is unique in the domain, whereas there is more than one planet). The two could be linked by a class assertion. The presence of such assertions and axioms allows an individual to “inherit” all the properties of the classes to which it belongs.

The concepts of a thesaurus and the classes of an ontology represent meaning in two fundamentally different ways. Thesauri express the meaning of a concept through terms, supported by adjuncts such as a hierarchy, associated concepts, qualifiers, scope notes and/or a precise definition, all directed mainly to human users. Ontologies, in contrast, convey the meaning of classes through machine-readable membership conditions.

In a thesaurus every concept has a preferred term to represent it (see ISO 25964-1:2011, 4.1), and often has one or more non-preferred terms to support comprehensive retrieval. Correspondingly, in an ontology, each class and individual usually (though not invariably) has a label. While it is possible to assign multiple labels to the same class, there is not generally any need for preferred nor non-preferred status since providing controlled access points for indexing and retrieval is not the primary objective of ontologies.

Disambiguation of natural language terms is more of an issue for thesauri than for ontologies. For example, a candidate term “moon” could have more than one meaning. In a thesaurus this ambiguity should be avoided, for example by using the plural form “moons” or “natural satellites” to apply to the moons of any planet, and

retaining the singular proper name “Moon” for the particular satellite that orbits the Earth. If both these concepts were included in one ontology, they would have different identifiers and so distinct term labels would not be essential.

As apparent in the example of the planets, the instance relationship used in some thesauri approximates to the class assertion used in ontologies. Likewise, the generic hierarchical relationship (BTG/NTG, which could be used in a thesaurus between “planets” and “celestial bodies”) corresponds to the subclass axiom in ontologies. However, in practice few thesauri make the distinction between generic, whole-part and instance relationships (see ISO 25964-1:2011, 10.2). The undifferentiated hierarchical relationship most commonly found in thesauri is inadequate for the reasoning functions of ontologies. Similarly the associative relationship is unsuited to an ontology, because it is used in a multitude of different situations (see ISO 25964-1:2011, 10.3.3) and therefore is not semantically precise enough to enable inferencing.

21.4 Interoperability with thesauri

21.4.1 Use cases

In other clauses of this part of ISO 25964, the main form of interoperability considered is mapping, with the aim of enabling conversion of search terms and/or indexing terms. In the case of ontologies, however, information retrieval applications that use ontologies directly for indexing are untypical and therefore mapping for this purpose will not be considered here. This clause instead considers the following three use cases, which are among the possibilities currently emerging:

- a) reengineering a thesaurus as an ontology;
- b) complementary use of a thesaurus and an ontology; and
- c) hybrid approaches.

21.4.2 Reengineering a thesaurus as an ontology

By exploiting the knowledge implicit in an existing thesaurus, reengineering aims to create an ontology that can be used for logical reasoning. In some cases, the reengineering may seek to merge a thesaurus with an existing formal upper ontology; the classes in the merged product being typically drawn from the original ontology at the upper levels and from the original thesaurus at lower levels.

To support reengineering a thesaurus as an ontology, a suitable formal language and other ontology tools should be chosen. The classes and individuals needed for the ontology should be selected from among the thesaurus concepts and (optionally) also supplemented by new classes and individuals. This often involves changing the hierarchical structure of the thesaurus. The concepts and relationships in the thesaurus should be examined, and reformulated in finer detail where necessary to eliminate any ambiguity prejudicial to the inferencing needs of the ontology. It might be necessary to distinguish the different thesaurus hierarchical relationships, if only the generalized BT/NT relationship is present. Additional axioms and other ontology features should be added as necessary.

The resulting reengineered ontology, accompanied by a set of assertions, could enable a range of new applications with inferencing capabilities. However, some of the qualities of the original thesaurus, designed to support indexers and searchers, might be lost in the process. The effectiveness of the reengineered product in retrieval cannot be taken for granted.

Before a reengineering project is undertaken, the feasibility of potential new applications involving logical reasoning should be examined. The likely cost/benefit should be considered; will the cost in time and resources of reengineering yield sufficient benefit in new applications or improved performance? In some cases, it might be possible to use an existing thesaurus in combination with an existing ontology, as described in the use cases below. The cost of reengineering a large thesaurus should not be underestimated.

21.4.3 Complementary use of a thesaurus and an ontology

Current interest in “Linked Data” has led to many projects (see, for example, the W3C Library Linked Data Incubator group report^[9]) that combine the reasoning power of ontologies with the retrieval capabilities of knowledge organization systems such as thesauri. (Guidance and use cases for the library user community are provided in Reference [9]). Examples of the following two scenarios are currently found, and in both of them Semantic Web standards such as SKOS^[37] and OWL^[35] are very often deployed.

- A thesaurus may be used to index items of a knowledge base, in order to support search and/or browsing of the knowledge base and its ontology. Meaningful reference information for the classes of the ontology can also be drawn from an accompanying set of documents indexed with the thesaurus. This type of approach supports a user whose starting point is natural language, rather than the formalized artificial language of the ontology.
- Complementary use of a thesaurus and an ontology may be mediated by a metadata element set (also known as a metadata schema). With this approach, the definitions and attributes of the metadata elements should be studied and a model of the application domain developed. An ontology should be developed, in which selected metadata elements are established as classes and/or properties. Typically, the range of certain metadata elements is constrained to the values in a particular controlled vocabulary, such as a thesaurus. This combination potentially allows logical inferencing at the level of the metadata elements, as well as supporting information retrieval via the terms and concepts of the controlled vocabularies.

EXAMPLE

The data model of the Europeana Data Model (EDM)^[20], draws on ontologies and metadata schemas such as ORE^[25], FOAF^[9], and Dublin Core^[17], while recommending the use of publicly accessible thesauri and other authority lists for the values to be used with specific metadata elements such as Subject, Creator, etc.

21.4.4 Hybrid approaches

It is also possible to use selected concepts and relationships from an existing thesaurus to derive a new ontology, without discarding the original thesaurus. In this scenario, the elements from one or more metadata schemas could also (optionally) be integrated into the ontology. Such a project involves both reengineering *and* complementary use of the resultant ontology. The insights gained from the development work could even feed back into improvement of the original vocabularies and schemas.

The emergence of the Semantic Web and associated ontology tools has stimulated a surge of research and development projects. Consequently this list of use cases is open-ended and new needs and opportunities for interoperability can be expected.

22 Terminologies

22.1 Key characteristics and background

22.1.1 General description

A terminology is a “set of designations belonging to one special language” (see ISO 1087-1:2000, 3.5.1), each designation representing a concept by means of a sign, symbol, term or appellation.

Terminological data can be presented in various formats, for example as term banks, termbases, glossaries or other publications.